DOCUMENT RESUME

ED 366 020 CS 508 427

AUTHOR Fowler, Carol A., Ed.
TITLE Speech Research Status Report, January-March 1993.
INSTITUTION Haskins Labs., New Haven, Conn.
REPORT NO SR-113
PUB DATE 93
NOTE 214p.; For the previous report, see ED 359 575.
PUB TYPE Collected Works - General (020) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC09 Plus Postage.
DESCRIPTORS Adults; *Articulation (Speech); *Beginning Reading; Chinese; Communication Research; Elementary Secondary Education; French; Higher Education; Infants; Language Acquisition; Language Research; Music; Reading Writing Relationship; *Speech Communication; Spelling; Thai
IDENTIFIERS Speech Research

ABSTRACT

One of a series of quarterly reports, this publication contains 14 articles which report the status and progress of studies on the nature of speech, instruments for its investigation, and practical applications. Articles in the publication are: "Some Assumptions about Speech and How They Changed" (Alvin M. Liberman); "On the Intonation of Sinusoidal Sentences: Contour and Pitch Height" (Robert E. Remez and Philip E. Rubin); "The Acquisition of Prosody: Evidence from French- and English-Learning Infants" (Andrea G. Levitt); "Dynamics and Articulatory Phonology" (Catherine P. Browman and Louis Goldstein); "Some Organizational Characteristics of Speech Movement Control" (Vincent L. Gracco); "The Quasi-Steady Approximation in Speech Production" (Richard S. McGowan); "Implementing a Genetic Algorithm to Recover Task-Dynamic Parameters of an Articulatory Speech Synthesizer" (Richard S. McGowan); "An MRI-Based Study of Pharyngeal Volume Contrasts in Akan" (Mark K. Tiede); "Thai" (M. R. Kalaya Tingsabadh and Arthur S. Abramson); "On the Relations between Learning to Spell and Learning to Read" (Donald Shankweiler and Eric Lundquist); "Word Superiority in Chinese" (Ignatious G. Mattingly and Yi Xu); "Prelexical and Postlexical Strategies in Reading: Evidence from a Deep and a Shallow Orthography" (Ram Frost); "Relational Invariance of Expressive Microstructure across Global Tempo Changes in Music Performance: An Exploratory Study" (Bruno H. Repp); and "A Review of 'Psycholinguistic Implications for Linguistic Relativity: A Case Study of Chinese' by Rumjahn Hoosain" (Yi Xu). (RS)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 366 020

Haskins Laboratories Status Report on Speech Research

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

SR-113
JANUARY-MARCH 1993

BEST COPY AVAILABLE

CS508427

Haskins
Laboratories
Status Report on

Speech Research

SR-113
JANUARY-MARCH 1993

NEW HAVEN, CONNECTICUT

Distribution Statement

Editor
Carol A. Fowler

Production
Yvonne Manning
Fawn Zefang Wang

This publication reports the status and progress of studies on the nature of speech, instrumentation for its investigation, and practical applications.

Distribution of this document is unlimited. The information in this document is available to the general public. Haskins Laboratories distributes it primarily for library use. Copies are available from the National Technical Information Service or the ERIC Document Reproduction Service. See the Appendix for order numbers of previous Status Reports.

Correspondence concerning this report should be addressed to the Editor at the address below:

Haskins Laboratories
270 Crown Street
New Haven, Connecticut
06511-6695

Phone: (203) 865-6163 *FAX:* (203) 865-8963 *Bitnet:* HASKINS@YALEHASK
Internet: HASKINS%YALEHASK@VENUS.YCC.YALE.EDU

This Report was reproduced on recycled paper

Acknowledgment

The research reported here was made possible in part by support from the following sources:

National Institute of Child Health and Human Development

Grant HD-01994
Grant HD-21888

National Institute of Health

Biomedical Research Support Grant RR-05596

National Science Foundation

Grant DBS-9112198

National Institute on Deafness and Other Communication Disorders

Grant DC 00121
Grant DC 00183
Grant DC 00403
Grant DC 00016
Grant DC 00594
Grant DC 00865
Grant DC 01147
Grant DC 00044
Grant DC 00825
Grant DC 01247

Investigators

Arthur Abramson*
Peter J. Alfonso*
Eric Bateson*
Fredericka Bell-Berti*
Catherine T. Best*
Susan Brady*
Catherine P. Browman
Claudia Carello*
Franklin S. Cooper*
Stephen Crain*
Lois G. Dreyer*
Alice Faber
Laurie B. Feldman*
Janet Fodor*
Anne Fowler*
Carol A. Fowler*
Louis Goldstein*
Carol Gracco
Vincent Gracco
Katherine S. Harris*
John Hogden
Leonard Katz*
Rena Arens Krakow*
Andrea G. Levitt*
Alvin M. Liberman*
Diane Lillo-Martin*
Leigh Lisker*
Anders Löfqvist
Ignatius G. Mattingly*
Nancy S. McGarr*
Richard S. McGowan
Patrick W. Nye
Kiyoshi Oshima†
Kenneth Pugh*
Lawrence J. Raphael*
Bruno H. Repp
Hyla Rubin*
Philip E. Rubin
Elliot Saltzman
Donald Shankweiler*
Jeffrey Shaw
Michael Studdert-Kennedy*
Michael T. Turvey*
Douglas Whalen

Technical Staff

Michael D'Angelo
Vincent Gulisano
Donald Hailey
Yvonne Manning
William P. Scully
Fawn Zefang Wang
Edward R. Wiley

Administrative Staff

Philip Chagnon
Alice Dadourian
Betty J. DeLise
Lisa Fresa
Joan Martinez

Students*

Melanie Campbell
Sandra Chiang
Margaret Hall Dunn
Terri Erwin
Joseph Kalinowski
Laura Koenig
Betty Kollia
Simon Levy
Salvatore Miranda
Maria Mody
Weijia Ni
Mira Peter
Christine Romano
Joaquin Romero
Dorothy Ross
Arlyne Russo
Michelle Sancier
Sonya Sheffert
Caroline Smith
Brenda Stone
Mark Tiede
Qi Wang
Yi Xu
Elizabeth Zsiga

*Part-time
†Visiting from University of Tokyo, Japan

Contents

Some Assumptions about Speech and How They Changed
Alvin M. Liberman 1

On the Intonation of Sinusoidal Sentences:
Contour and Pitch Height
Robert E. Remez and Philip E. Rubin 33

The Acquisition of Prosody: Evidence from French- and
English-Learning Infants
Andrea G. Levitt 41

Dynamics and Articulatory Phonology
Catherine P. Browman and Louis Goldstein 51

Some Organizational Characteristics of Speech
Movement Control
Vincent L. Gracco 63

The Quasi-steady Approximation in Speech Production
Richard S. McGowan 91

Implementing a Genetic Algorithm to Recover Task-dynamic
Parameters of an Articulatory Speech Synthesizer
Richard S. McGowan 95

An MRI-based Study of Pharyngeal Volume Contrasts in Akan
Mark K. Tiede 107

Thai
M. R. Kalaya Tingsabadh and Arthur S. Abramson 131

On the Relations between Learning to Spell and
Learning to Read
Donald Shankweiler and Eric Lundquist 135

Word Superiority in Chinese
Ignatius G. Mattingly and Yi Xu 145

Prelexical and Postlexical Strategies in Reading:
Evidence from a Deep and a Shallow Orthography
Ram Frost 153

Relational Invariance of Expressive Microstructure across
Global Tempo Changes in Music Performance:
An Exploratory Study
Bruno H. Repp 171

A Review of *Psycholinguistic Implications for Linguistic Relativity:
A Case Study of Chinese* by Rumjahn Hoosain
Yi Xu 197

Appendix 209

Haskins
Laboratories
Status Report on

Speech Research

Haskins Laboratories Status Report on Speech Research
1993, SR-113, 1-32

Some Assumptions about Speech and How They Changed*

Alvin M. Liberman

My aim is to provide a brief account of the research on speech at Haskins Laboratories as seen from my point of view. In pursuit of that aim, I will scant most experiments and their outcomes, putting the emphasis, rather, on the changing assumptions that, as I understood them, guided the research or were guided by it. I will be particularly concerned to describe the development of those assumptions, hewing as closely as I can to the order in which they were made, and then either abandoned or extended.

My account is necessarily inaccurate, not just because I must rely in part on memory about my state of mind almost 50 years ago when, in the earliest stages of our work, we did not always make our underlying assumptions explicit, but, even more, for reasons that put a proper account beyond the reach of any recall, however true. The chief difficulty is in the relation between the theoretical assumptions and the research they were supposed to rationalize. Thus, it happened only once, as I see it now, that the assumptions changed promptly in response to the results of a particular experiment; in all other cases, they lagged behind, as reinterpretations of data that had accumulated over a considerable period of time. Moreover, theory was influenced. not only by our empirical findings, but equally, if even more belatedly, by general considerations of plausibility that arose when, stimulated by colleagues and by the gradual broadening of my own outlook, I began to give proper consideration to the special requirements of phonological communication and to all that is entailed by the fact that speech is a species-typical product of biological evolution. The consequence for development of theory was that, with the one exception just noted, I never changed my mind abruptly, but rather segued from each earlier position to each later one, leaving behind no clear sign to mark the time, the extent of the change, or the reason for making it. Faced now with having to describe a theory that was in almost constant flux, I can only offer what I must, in retrospect, make to seem a progress through distinct stages.

My account is also necessarily presumptuous, because I will sometimes say 'we' when I probably should have said 'I', and vice versa. In so doing, I am not trying to avoid blame for bad ideas that were entirely mine, nor to claim credit for good ideas that I had no part in hatching. It is, rather, that everything I have done or thought in research on speech has been profoundly influenced by my colleagues at the Laboratories. In most cases, however, I can't say with certainty just how or by whom. A consequence is that, of all the words I use in this chronicle, the most ambiguous by far are 'I' and 'we'.

'I' began to be confused with 'we' on a day in June, 1944, when I was offered a job at the Laboratories to work with Frank Cooper on the development of a reading machine for the blind, a device that would convert printed letters into intelligible sounds. Of course, we appreciated from the outset that the ideal machine would render the print as speech that the blind user had already mastered. Though that is done quite routinely now, it was, in 1944, far beyond the science and technology that was available to us. There were no optical character readers to identify the letters, and no rules for synthesis that would have enabled us to produce speech from their outputs. But, reassured by our assumptions about the relation of speech to language, of which more

The preparation of this paper was aided by the National Institute of Child Health and Human Development under Grant HD 01994. Since the paper is a very personal account of the research that has, since 1965, been supported by that agency, I am moved here to offer special thanks to the agency for its long-continued help, and to Dr. James Kavanagh, a member of its staff, for the kind and wise counsel that he has, for so many years, given me.

later, we did not think it critical that the machine should speak. Rather, we supposed that it had only to produce distinctive sounds of some kind that the blind would then learn to associate with the consonants and vowels of the language, much as we supposed they had done at an earlier stage of their lives with the sounds of speech. Thus conceived, our enterprise lay in the domains of auditory perception and discrimination learning, two subjects I had presumably mastered as a graduate student at Yale. Indeed, my dissertation had been based on experiments about discrimination learning to acoustic stimuli, and I was thoroughly familiar with the neobehaviorist, stimulus-response theory that was thought by my professors and me to account for the results. In the course of my education in psychology, I had learned nothing about speech, but I didn't think that mattered, because the theory I had absorbed was supposed, like most other theories in psychology, to apply to everything. I was, therefore, enthusiastic about the job, confident that I knew exactly how to make the reading machine serve its intended purpose and so put the theory to practical use. In the event, the theory, and, indeed, virtually everything else I had learned, proved to be, at best, irrelevant and, at worst, misleading. I think it unlikely that I would ever have discovered that had it not been for two fortunate, as they proved to be, circumstances. One was the collaboration, from the outset with Frank Cooper, whose gentle but nonetheless insistent prodding helped me to accept that the theory might be wrong, and to see what might be more nearly right. The other was that speech lay constantly before me, providing an existence proof that language could be conveyed efficiently by sound, and thus setting the high standard by which I was bound to evaluate the performance of the acoustic substitutes for speech that our early assumptions led us to contrive. But for Frank Coooper, on the one hand, and speech on the other, I might still be massaging those substitutes and modifying the conditions of learning, satisfied to achieve trivial improvements in systems that were, by comparison with speech, hopelessly inadequate. As it was, experience in trying to find an alternative set of sounds brought Frank and, ultimately, me to the conclusion that speech is uniquely effective as an acoustic vehicle for language. It remained only to find out why.

But I get ahead of the story. To set out from the proper beginning, I should say why we initially believed there was nothing special about speech or its underlying processes, and where that belief led us, not only in the several stages of the reading machine work, but also in the development of, and early work with, a research synthesizer we called the Pattern Playback.

The assumptions about speech that have been made by us and others differ in many details. As I see it now, however, there is one question that stands above those details, dividing theories neatly into two categories: does speech recruit motor and perceptual processes of a general sort, processes that cut horizontally across a wide variety of behaviors; or does it belong to a special phonetic mode, a distinct kind of action and perception comprising a vertical arrangement of structures and mechanisms specifically and exclusively adapted for linguistic communication? I will use this issue as the basis for organizing my account of the assumptions we made, calling them 'horizontal' or 'vertical' according to the side of the divide on which they fall.

THE HORIZONTAL VIEW OF SPEECH AND THE DESIGN OF READING MACHINES

As it pertains to the perceptual side of the speech process, the horizontal view, which is how we saw speech at the outset, rests on three assumptions: (1) The constituents of speech are sounds. (2) Perception of these sounds is managed by processes of a general auditory sort, processes that evoke percepts no different in kind from those produced in response to other sounds. (3) The percepts evoked by the sounds of speech, being inherently auditory, must be invested with phonetic significance, and that can be done only by means of a cognitive translation. Accordingly, the horizontal view assumes a second stage, beyond perception, where the purely auditory percepts are given phonetic names, measured against phonetic prototypes, or associated with 'distinctive features'. Seen this way, perceiving speech is no different in principle from reading script; listener and reader alike perceive one thing and learn to call it something else.

These assumptions were—and, perhaps, still are—so much a part of the received view that we could hardly imagine an alternative; it simply did not occur to us to question them, or even to make them explicit. At all events, it was surely these conventional assumptions that gave us confidence in our assumption that nonspeech sounds could be made to work as well as speech, for what they clearly implied was that the processes by which blind users would learn to 'read' our sounds would

differ in no important way from those by which they had presumably learned to perceive speech. Our task, then, was simply to contrive sounds of sufficient distinctiveness, and then design the procedures by which the blind would learn to associate them with language.

Auditory discriminability is all it takes

As for distinctiveness, I was, at the outset, firmly in the grip of the notion that it was just so much psychophysical discriminability, almost as if it were to be had simply by separating the stimuli one from another by a sufficient number of just-noticeable-differences. This notion fit all too nicely with our ability, even then, to engineer a print-to-sound machine that would meet the discriminability requirement, for such a machine had only to control selected parameters of its acoustic output according to what was seen by a photocell positioned behind a narrow scanning slit. Provided, then, that we chose wisely among the many possible ways in which the sound could be made to vary according to what the photocell saw, the auditory results would be discriminable, hence learnable.

In all these machines the sound would be controlled directly by the amount or vertical position of the print under the scanning slit. We therefore called them 'direct translators' to distinguish them from an imaginable kind we called 'recognition machines', having in mind a device that might one day be available to identify each letter, as present-day optical character readers do, and produce in response a preset sound of any conceivable type. Since we were in no position to build a recognition machine, we set our sights initially on a direct translator.

We were aware that a reading machine of the direct-translation kind had been constructed and tested in England just after World War I. This machine, called the Optophone, scanned the print with five beams of light, each modulated at a different audio frequency. When a beam encountered print, a tone was made to sound at a frequency (hence, pitch) equal to the frequency at which that beam was modulated. Thus, as the scanning slit and its beams were moved across the print, the user would hear a continuously changing musical chord, the composition of which depended, instant by instant, on the modulation frequencies of the beams that struck print. The user's task, of course, was to learn to interpret the changing chords as letters and words. I recall not ever knowing what tests, if any, had been carried out to determine just how useful the machine was, only that a blind woman in Scotland had been able to make her way through simple newspaper text. We did know that the machine was not in use anywhere in 1944, so we could only assume that it had been found wanting. At all events, the lesson I took from the Optophone was not that an acoustic alphabet is no subtitute for speech, but that the particular alphabet produced by the Optophone was not a good one. Moreover, it seemed likely, given the early date at which the machine had been made, that the signal was less than ideally clear. I, perhaps more than Frank, was sure we could do better. (Indeed, I think that Frank, even at this early stage, had reservations about any machine that read letter by letter, but he shared my belief that we could design a letter-reading machine good enough to be of some use.)

Doing better, as we conceived it then, simply meant producing more learnable sounds. Unfortunately, there was nothing in the literature on auditory perception to tell us how to do that, so we set out to explore various possibilities. For that purpose, Frank built a device with which we could simulate the performance of almost any direct translator, with the one limitation that the subjects of our experiments were not free to choose the texts or to control the rate of scan. (This was because the print had to be scanned by passing a film negative of the text across a fixed slit.) Otherwise, the simulator was quite adequate for producing the outputs we would want to screen before settling on those that deserved further testing.

For much of the initial screening, Frank and I made the evaluations ourselves after determining roughly how well we could discriminate the sounds for a selected set of letters and words. On this basis, we quickly came to the conclusion that there was, perhaps, some promise in a system that used the print to modulate the frequency of the output signal. (Certainly, this was very much better than modulating the amplitude, which had been tried first.) Having made this early decision, we experimented with different orientations and widths of the slit, the shape of the sound wave, and the range of modulation, selecting, finally, a vertical orientation of a slit that was somewhat narrower than, say, the vertical bar of a lower case 't', and a sine wave with a modulation range of 100 to 4000 Hz.

Further informal tests with this frequency-modulation (FM) system showed that it failed to provide a basis for discriminating certain letters (for example, n and u), which led us to conclude that it would likely prove in more thorough tests

to be not good enough. This conclusion did not discourage me, however, for I reckoned that the difficulty was only that the signal variation, being one-dimensional, was not sufficiently complex. Make the signals more complexly multi-dimensional, I reasoned, so that somewhere in the complexity the listener would find what he needed as the basis for associating that signal with the letter it represented. I recall that here, too, Frank was skeptical, but, open minded as always, he agreed to try more elaborate arrangements. In one, the Dual FM, the slit was divided into an upper and lower half (each with its own photocell), so that two tones, one in the upper part of the 4000-Hz range, the other in the lower, would be independently modulated. In another (Super FM), the slit was divided into thirds (each with its own photocell). and the difference between the output of the middle cell and the sum of the upper and lower cells controlled the (relatively slow) frequency modulations of a base tone, while the sum of all three cells controlled an audio frequency at which the base tone was itself frequency modulated. The effect of this latter modulation was to create side bands on the base frequency at intervals equal to the modulation frequency, and thus cause frequent, sudden, and usually gross changes in timbre and pitch as a consequence of the amount and position of the print under the slit. At all events, the signals of the Super FM seemed to me to vary in wonderfully complex ways, and so to provide a fair test of my assumption about the necessary condition for distinctiveness and learnability. We also simulated the Optophone, together with a variant on it in which, in a crude attempt to create consonant-like effects, we had the risers and descenders of the print produce various hisses and clicks. We even tried an arrangement in which the print controlled the frequency positions of two formant-like bands, giving an impression of constantly changing vowel color.

We tested all of these systems, together with three or four others of similar type, by finding how readily subjects could learn to associate the sounds with eight of the most common four-letter words. To determine how well these systems did in relation to speech, we applied the same test to an artificial language that we spoke simply by transposing phonological segments. Thus, vowels were converted into vowels, stops into stops, fricatives into fricatives, etc., so the syllable structure remained unchanged, hence easy for our human reading machine to pronounce. We called this new language 'Wuhzi', in honor of one of the transposed words.

The results of the learning test were simple enough: of all the nonspeech signals tested, the original, simple FM and the original Optophone were the most nearly learnable; all others trailed far behind. Worst of all, and by a considerable margin, were the extremely complex signals of the Super FM, the system for which I had entertained such high hopes. As for Wuhzi, the transposed speech, it was in a class of its own. It was mastered (for the purposes of the test) in about 15 trials, in contrast to the nonspeech sounds, on which subjects were, after 24 trials, performing at a level of only 60% with the simple FM and Optophone signals, and at 50% or lower for the others, and all seemed at, or close to, their asymptotes. More troubling was the fact that learnability of the nonspeech signals went down appreciably as the rate of scan was increased, even at modest levels of rate. What should have been most troubling, however, were the indications that learning would be rate specific. Though we did not pursue the point, it seemed clear in my own experience with the FM system that what I had learned to 'read' at a rate of, say, 10 words per minute, did not transfer to a rate of 50. My impression, as I remember it now, was not that things sounded much the same, only speeded up, but rather that the percept had changed entirely, with the result that I could no longer recognize whatever I had learned to respond to at the slower speed. To the extent that this observation was correct, users would have had to learn a different nonspeech 'language' for every significantly different rate of reading, and that alone would have made the system impractical. It also became clear that at rates above 50 or so words per minute, listeners could identify the words only on the basis of some overall impression—I am reluctant to call it a pattern—in which the individual components were not readily retrievable. The consequence of this, though we did not take proper account of it at the time, was that perception of the nonspeech signals would have been, at best, logophonic, as it were, not phonologic, so users would have had to learn to identify as many sounds are there were words to be read, and would not have been able to fall back on their phonologic resources in order to cope with new words.

At this point it had become evident that, as Frank had been saying for some time, perception of the acoustic alphabet produced by any letter-

reading machine would be severely limited by the temporal resolving power of the ear—that is, by its poor abililty to segregate and properly order acoustic events that are presented briefly and in rapid succession. Taking account of the number of discrete frequency peaks in each alphabetic sound and the average number of alphabetic characters per word, we estimated the maximum rate to be around 50 words per minute, which I now believe to have been very much on the high side of what might have been achieved. Still, 50 words per minute would be quite useful for many purposes, and we had a considerable investment in the letter-reading direct-translator kind of system, so we felt obliged to see how far a subject could go, given control of the scanning rate, a great variety of printed material, and much practice. Frank therefore built a working model of a direct-translating FM reading machine, and we set several subjects to work trying to master it. After 90 hours of practice, our best subject was able to manage a fifth-grade reader (without correction) at an average speed of about four words per minute. Moreover, the subject seemed to have reached her peak performance during the first 45 hours. Comparing performance at the end of that period with performance after 90 hours, we found no increase in rate and no decrease in errors. Tests of the same system conducted by John Flynn at the Naval Medical Research Laboratory yielded results that gave no more grounds for optimism.

I must not omit from this account a reference to our experience with a sighted subject named (dare I say, appropriately?) Eve, who, having first produced a properly shaped learning curve, attained a reading speed of over 100 words per minute, and convincingly demonstrated her skill before the distinguished scientists who were sponsoring our research. In that demonstration, as in all of her work with the machine, she wore blacked-out welder's goggles so she would not suffer the discomfort of having to keep her eyes shut tight. We all remember the day when one of our number, acting on what seemed an ungentlemanly impulse, put an opaque screen between her and the print she was 'reading', at which point she bounded from her chair and fled the lab, confessing later to the young man who had been monitoring her practice sessions that, by leaning her cheek against her fist, she had, from the very beginning, been raising the bottom of the goggles and so seeing the print. In designing the training and testing procedures, I had controlled for every possibility except that one. Obviously, I was the wrong kind of psychologist.

Having concluded at last that, given the properties of the auditory system, letter-by-letter reading machines were destined to be unsatisfactory, we experimented briefly with a system in which the sound was controlled by information that had been integrated across several letters. This was intended to reduce the rate at which discrete acoustic events were delivered to the ear, and so circumvent the limitation inposed by its temporal resolving power. I cannot now imagine what we thought about the consequences of having to associate holistically different sounds with linguistically arbitrary conjunctions of letters. But whatever it was that we did or did not have in mind, this integrating type of reading machine failed every test and was quickly abandoned.

One other scheme we tested deserves mention if only because it now seems incredible that we should ever have considered it. I suppose we felt obliged to look forward to the time when optical character readers would be available, and so to test some presumably appropriate output of the 'recognition' machine that would then be possible. Prerecorded 'phonemes' seemed an obvious choice. After all, the acoustic manifestations of phonemes were known to be distinctive, and people had, on the horizontal view, already learned to connect them to language. As for difficulty with rate, we must have supposed that these sounds carried within them the seeds of their own integration into larger wholes. At all events, we recorded various 'phonemes' on film—'duh' for /d/, 'luh' for /l/, etc.—and then carefully spliced the pieces of film together into words and sentences. The result was wholly unintelligible, even at moderate rates, and gave every indication of forever remaining so.

Perhaps auditory patterning is the answer

From all this experience with nonspeech we drew two conclusions, one right, one wrong. The right conclusion was that an acoustic alphabet is no way to convey language. Of course, it was humbling for me to realize that I might have reached that conclusion without all the work on the reading machines, simply by measuring the known temporal-resolving characteristics of the ear against the rate at which the discrete sounds would have to be perceived. As I have already suggested, Frank must have thought this through, though he might not have been sufficiently pessimistic about the limits, and that would account for his early opinion that an acoustic alphabet might be at least marginally useful. I, on the other hand, took the point only after seeing our early results and having my nose rubbed in them.

The wrong conclusion—wrong because it was still uncompromisingly horizontal in outlook—was that what we needed was not discriminability but proper auditory patterning. Speech succeeds, we decided, because its sounds conform to certain Gestalt-like principles that yield perceptually distinctive patterns, and cause the phonetic elements to be integrated into the larger wholes of syllables and words. Apparently, it did not occur to me that phonological communication would be impossible if the discrete and commutable elements it requires were to be absorbed into indissoluble Gestalten. Nor did I bother to wonder how speakers might have managed their articulators so as to force all of the many covarying but not-independently-controllable aspects of the acoustic signal to collaborate in such a way as to produce just those sounds that would form distinctive wholes.

In any case, nobody knew what the principles of auditory pattern perception were. Research in audition, unlike the corresponding effort in vision, had been quite narrowly psychophysical, having been carried out with experimental stimuli that were not sufficiently complex to reveal such pattern perception as there might be. So if, for the purposes of the reading machine, we were going to make its sounds conform to the principles that underlie good auditory patterns, we had first to find the principles. We therefore decided to turn away from the reading machine work until we should have succeeded in that search. We were the more motivated to undertake it, because the outcome would not only advance the construction of a reading machine, but also count, more generally, as a valuable contribution to the science of auditory perception.

THE PATTERN PLAYBACK: MAKING THE RIGHT MOVE FOR THE WRONG REASON

It is exactly true, and important for understanding how horizontal our attitude remained, to say that our aim at this stage was to study the perception of patterned sounds in general, not speech in particular. As for the reading machine, we did not suppose we would use our results to make it speak, only that we would have it produce sounds that would be as good as speech because they would conform to the same principles of auditory perception. These would be sounds that, except for the accidents of language development, might as well have been speech. Obviously, we would look to speech for examples of demonstrably well-patterned sounds. Indeed, we would be especially attentive to speech, but only because we did not know where else to search for clues. At this stage, however, our interest lay in auditory perception.

Since we barely knew where to begin, we expected to rely, especially at the outset, on trial and error. Accordingly, we had to have some easy way to produce and modify a large number and wide variety of acoustic patterns. The sound spectrograph and the spectrograms it produced were no longer a wartime secret, and, like many others, we were impressed with the extent to which spectrograms made sense to the eye. Beyond a doubt, they provided an excellent basis for examining speech; it seemed to us a small step to suppose that they would serve equally well for the purpose of experimenting with it. The experimenter would have immediately available the patterned essence of the sound, and thus easily grasp the parameters that deserved his attention. These would be changed to suit his aims, and he would then have only to observe the effects of the experimental changes on the sound as heard. But that required a complement to the spectrograph—that is, a device that would convert spectrograms, as well as the changes made on them, into sound.

It was in response to these aims and needs that the Pattern Playback was designed and, in the case of its most important components, actually built by Frank Cooper. My role was simply to offer occasional words of encouragement, and, as the device took shape, to appreciate it as a triumph of design, a successful realization of Frank's determination to provide experimental and conceptual convenience to the researcher. The need for that convenience is, I think, hard to appreciate fully, except as one has lived through the early stages of our speech research, when little was known about acoustic phonetics, and progress depended critically, therefore, on the ability to test guesses at the rate of dozens per day.

Seven years elapsed between the start of our enterprise and the publication of the first paper for which we could claim significance. I would therefore here record that the support of the Haskins Laboratories and the Carnegie Corporation of New York was a notable. and, to Frank and me, essential exercise of faith and patience, Few universities or granting agencies would then have been, or would now be, similarly supportive of two young investigators who were trying for such a long time to develop an unproved method to investigate a question—what is special about speech that it works so well?—that was not otherwise being asked. I managed to survive in academe by changing universities frequently, thus

confusing the promotion-tenure committees, and also by loudly trumpeting two rat-learning studies that I published as extensions of my thesis research.

There were many reasons it took so long to get from aim to goal. One was the failure, after two years of work, of the first model of the device that was to serve as our primary research tool. It was the second and very different model (hereafter the Pattern Playback or simply the Playback) that was, for our purposes, a success. In the form in which it was used for so many years, this device captured the information on the spectrogram by using it to control the fate of frequency-modulated light beams. Arrayed horizontally so as to correspond to the frequency scale of the spectrogram, there were 50 such modulated beams, comprising the first 50 harmonics of a 120 Hz fundamental. Those beams that were selected by the pattern of the spectrogram were led to a phototube, the resulting variations of which were amplified and transduced to sound. In effect, the device was, as someone once said, an optical player piano.

When the Pattern Playback was conceived, we thought we might not be able to work from hand-painted copies of spectrograms, but only from real ones. Therefore, we needed spectrograms on film so that, given photographic negatives, the phototube would see the beams that passed through the film. (When operating from hand-painted patterns, the phototube would receive the beams that were reflected from the paint.) For convenience in manipulating the patterns, we also wanted the frequency and time scales to be of generous dimensions. Moreover, we thought at this stage that we would need spectrograms with a large dynamic range. None of these needs was fully met by the spectrograph that had been built originally at the Bell Telephone Laboratories, and it was not available for sale in any case, so Frank set about to design and build our own. Unfortunately for the progress of our research, that took time. As for the Playback itself, it was, as our insurance inspector remarked when first he saw it, 'homemade'. And, indeed it was. Only the raw materials were store-bought. Everything else was designed and fashioned at the Laboratories, including, especially, the tone wheel, the huge circular film with 50 concentric rings of variable density used to produce the light beams that played on the spectrograms.

Once the special-purpose spectrograph and the Playback were, at last, in working order, we had to determine that speech could survive the rigors of the transformations wrought first by the one machine and then by the other. I well remember the relief I felt—Frank probably had more faith and therefore experienced correspondingly less relief--when, operating from a negative of a spectrogram, the Playback produced, on its first try, a highly intelligible sentence.

But that was only to pass the first test. For using these 'real' spectrograms as a basis for experimenting with speech would have been awkward in the extreme, since it would have been very hard to make changes on the film. We therefore began immediately to develop the ability to use hand-painted spectrograms to control the sound. For that purpose, we had first to find a paint that would wet the acetate tape (i.e., not bead up), be easily erased (without leaving scars), and dry quickly. (The last requirement became nearly irrelevant when, quite early in the work, we acquired a hair-dryer.) Unable to find a commercially available paint that met our specifications, we became paint manufacturers, making our way by trial and error to a satisfactory product. That much we had to do. But there was time spent in other preliminaries that proved to be quite unnecessary. Thus, overestimating the degree of fidelilty to the original spectrogram we would need, we assumed we would have to control the relative intensities of the several parts of the pattern, and so devoted considerable effort to preparing and calibrating paints of various reflectances. We also supposed that we would have to produce gradations of a kind that could best be done with an airbrush, so we fiddled for a time with that.

But then, having decided that the price of fidelity was too high, we simply began, with an artist's brush and our most highly reflecting paint, to make copies of spectrographic representations of sentences, using for this purpose a set of twenty that had been developed at Harvard for work on speech intelligibility. Taking pains with our first copies to preserve as much detail as possible, we succeeded in producing patterns that yielded what seemed to us a level of intelligibility high enough to make the method useful for research. We were then further encouraged about the prospects of the method when, after much trial and error, we discovered that we could achieve even greater intelligibility—about 85% for the twenty sentences—with still simpler and more highly schematized patterns.

Having got this far, we ran a few preliminary experiments no different in principle from those, very common at the time, in which intelligibility was measured as a function of the filtering that the speech signal had been subjected to. But

instead of filtering, we presented the formants one at a time and in all combinations. The grossly statistical result seemed of no great interest, even then, so we did not pursue it. There was, however, one unintended and unhappy result of our interest in the relative contributions of the several formants, In the course of giving a talk about the Playback at a meeting of the Acoustical Society of America, Frank played several copy-synthetic sentences, first with each formant separately, and then with the formants in various combinations. It was plain from the reaction of the audience that everyone was greatly surprised each time Frank read out the correct sentence. Apparently, people had formed wrong hypotheses as they tried to make sense of the speech produced by the individual formants, and subsequently had trouble correcting those hypotheses when presented with the full pattern. So the sentences got nowhere near the 85% intelligibility they deserved, and Frank got little recognition on that occasion for a promising research method.

AN EXCURSION INTO NONSPEECH AND THE INFINITELY HORIZONTAL HYPOTHESIS

We were, at this stage, strongly attracted to the notion that spectrograms of speech were highly readable because, as tranformations of well-patterned acoustic signals, they managed still to conform to basic principles of pattern perception. Implicit in this notion was the assumption that principles of pattern perception were so general as to hold across modalities, provided proper account was taken of the way coordinates were to be transformed in moving from the one modality to the other. As applied to the relation between vision and audition, this assumption could be tested by the extent to which patterns that looked alike could be so transformed as to make them sound alike. To apply this test to the spectrographic transform, we varied the size and orientation of certain easily identifiable geometric forms, such as circles, squares, triangles, and the like, converted them to sound on the Playback, and then asked whether listeners would categorize the sounds as they had the visual patterns. Under certain tightly constrained circumstances, the answer was yes, they would; otherwise, it was no, they would not. At all events, we were so bold as to publish the idea, together with a description of the Playback in the Proceedings of the National Academy of Sciences. Fortunately for us, that journal is not widely read in our field, so our brief affair with the mother of all horizontal views has remained till now a well-kept secret.

We also experimented briefly with a phenomenon that would, today, be considered an instance of 'streaming', though I thought of it then as the auditory equivalent of the phi phenomenon that had for so long occupied a prominent place in the study of visual patterns. Using painted lines so thin that each one activated only a single harmonic, we observed that when they alternated between a lower and a higher frequency, the listener heard the alternation only if the resulting sinusoids were sufficiently close in frequency and sufficiently long in duration; otherwise, the impression was of two streams of tones that bore no clear temporal relation to each other. In pursuing this effect, we looked for that threshold of frequency separation and duration at which the subject could not distinguish a physically alternating pattern from one in which the two streams of tones came on and went off in perfect synchrony. We did, in fact, succeed in finding such thresholds and in observing that they were sensitive to frequency separation and duration, as our hypothesis predicted, but, after exhibiting a certain threshold for a while, the typical subject would quite suddenly begin to hear the difference and to settle down, if only temporarily again, at a new threshold. Discouraged by this apparent lability, we abandoned the project and began to put our whole effort on speech.

EARLY (AND STILL HORIZONTALLY ORIENTED) RESEARCH ON SPEECH

It was at about this point that Pierre Delattre came to the lab to visit for a day and, fortunately for us, stayed for ten years. Initially, his interest was in discovering the acoustic basis for nasality in French vowels, but once he found what the Playback was capable of, his ambition broadened to include any and all elements of phonetic structure. So, while continuing to work on nasality, Pierre applied himself (and us) to producing two-formant steady-state approximations to the cardinal vowels of Daniel Jones. (Pierre supplied the criterial ear.) The result was published in Le Maitre Phonetique, written in French and in a narrow phonetic transcription. Thus, in our second paper, we continued the habit that had been established in the first of so publishing as to guarantee that few among our colleagues would read what we wrote.

Meanwhile, we had begun to put our attention once again on the copy-synthetic sentences, but now, instead of inquiring into the contribution of

each formant to overall intelligibility, we sought, more narrowly, to find the acoustic bases—the cues—for the perception of individual phones. I recall working first on the word 'kill' as it appeared in simplified copy we had made from a spectrogram of the utterance, 'Never kill a snake with your bare hands.' Looking for the phone [l], we held the pattern stationary at closely spaced temporal intervals, thus creating at each resting point a steady-state signal controlled by the static positions of the formants at that instant. The result, to our ears, was a succession of vowel-like sounds, but nothing we would have called an [l]. Yet, running the tape through at normal speed, and thus capturing the movement of the formant, produced a reasonable approximation to that phone. I don't remember what we made of this (to me) first indication of the importance of the dynamic aspects of the signal, but we did not immediately pursue it, turning instead to the [k] at the beginning of the same word.

Context-conditioned variability and the horizontal version of the motor theory

The advantage of [k] from our point of view as experimenters was that it appeared, in the copy-synthetic word 'kill', to be carried by a clearly identifiable and nicely delimited cue: a brief burst of sound that stood apart from the formants. Since we knew that [k] was a voiceless stop, in the same class as [p] and [t], we reckoned that all three stops might depend on such a burst, according to its position on the frequency scale. So, after a little trial and error, we carried out what must be counted our first proper experiment. Bursts (about 360 Hz in height, 15 msec in duration at their widest, and of a shape that caused Pierre to call them 'blimps') were centered at each of 12 frequencies covering the full range of our spectrograms. Each burst was placed in front of seven of Pierre's steady-state cardinal vowels, and the resulting stimuli were randomized for presentation to listeners who would be asked to identify the consonant in each case as [p], [t], or [k].

Of all the synthetic patterns ever used, these burst-plus-steady-state-vowel 'syllables' were undoubtedly the farthest from readily recognizable speech. Indeed, they so grossly offended Pierre's phonetic sensibilities that he agreed only reluctantly to join Frank and me as a subject in the experiment. After discharging my own duty as a subject, I thought, as did Frank, that Pierre's reservations were well taken, and that our responses would reveal little about stop consonants or, indeed, anything else. We were therefore surprised and pleased when, on tabulating the results, we saw a reasonably clear and systematic pattern, and then equally surprised and pleased when a group of undergraduates, known for technical purposes as 'phonetically naive subjects', did much as we had done, if just a little more variably. For us and for them, the modal [k] response was for bursts at or slightly above the second formant of the vowel, wherever that was; [t] was assigned most often to bursts centering at frequencies above the highest of the [k] bursts; and a [p] response was given to most of the bursts not identified as [t] or [k].

This experiment was, for me, an epiphany. It is the one, referred to earlier, that changed my thinking within hours or days after its results were in. What caused the change was the finding that the effect of an acoustic cue depended to such a very large extent on the context in which it appeared, and, more to the point, that perception accorded better with articulatory gesture than with sound.

The effect of context was especially apparent in the fact that the burst most frequently perceived as [k] was the one lying at or slightly above the second formant of the following vowel, even though that formant sampled a range that extended from 3000 Hz (for [i]) at the one end, to 700 Hz (for [u]), at the other. Moreover, this evidence that the same phonetic percept was cued by stimuli that were acoustically very different was but one side of the coin, for the results also showed that, given the right context, different percepts could be evoked by stimuli that were acoustically identical. This was the case with the burst at 1440 Hz, which was perceived predominantly as [p] before [i], as [k] before [a], and then again as [p], though weakly, before [u].

As an empirical matter, then, this first, very crude experiment demonstrated a kind and degree of context-conditioned variability in the acoustic cues that subsequent research has shown to be pervasive in speech. As for its relevance to our earlier work on reading machines, it helped to rationalize one of the conclusions we had been brought to, which was that speech cannot be an acoustic alphabet; for what the context effects showed was that the commutable *acoustic* unit is not a phone, but rather something more like a syllable.

From a theoretical point of view, the results revealed the need to find, somewhere in the perceptual process, an invariant to correspond to the invariant phonetic unit, and they strongly suggested

that the invariant is in the articulation of the phone. Thus, in the case of the [k] burst we noted that it was the articulatory movement—raising and lowering the tongue body at the velum—that remained reasonably constant, regardless of the vowel, and, further, that coarticulation of the constant consonant with the variable vowels accounted for the extreme variability in the acoustic signal. As for the other side of the coin—the very different perception of the burst before different vowels—which resulted, it should be noted, from a fortuitous combination of circumstances probably unique to this experiment, we supposed that, in order to produce something like a burst at 1440 Hz in front of [i] or [u], one had to close and open at the lips, while in front of [a], closing and opening at the velum was required.

Taking all this into account, we adopted a notion—the Early Motor Theory—that I now believe to have been partly right and partly wrong. It was right, I think, in assuming that the object of perception in phonetic communication is to be found in the processes of articulation. It was wrong, or at least short of being thoroughly right, because it implied a continuing adherence to the horizontal view. As earlier indicated, that view assumes a two-stage process: first an auditory representation no different in kind from any other in that modality, followed, then, by linkage to something phonetic, presumably as a result of long experience and associative learning. The phonetic thing the auditory representation becomes associated with is, in the conventional view, a name, prototype, or distinctive feature. As we described the Early Motor Theory in our first papers, we made no such explicit separation into two stages, but only because our concern was rather to emphasize that perception was more closely associated with articulatory processes than with sound, and then to infer that this was because the listener was responding to the sensory feedback from the movements of the articulators. However, we offered no hint that phonetic perception takes place in a distinct modality, thus omitting to make explicit the assumption that is, as will be seen, the heart of the vertical view. Indeed, we implied, to the contrary, that the effects we were concerned with were, potentially at least, perfectly general, and so would presumably occur for any perceptual response to a physical stimulus, given long association between the stimulus and some particular muscular movement, together with its sensory feedback. What was special about speech was only that it provided the par excellence example of the opportunity for precisely that association to be formed. In any case, my own view, as I remember it now, did comprehend two more or less distinct stages: an initial auditory representation that, as a result of associative learning, ultimately gave way to the sensory consequences of the articulatory gesture that had always been coincident with it. I had, I must now suppose, not spent much time wondering exactly what 'gave way' might mean. Had I been challenged on this point, I think I should have said that in the early stages of learning there surely was a proper auditory percept, no different in kind from the response to any other acoustic stimulus and equally available to consciousness, but that later, when the bond with articulation was secure, this representation would simply have ceased to be part of the perceptual process. But however I might have responded, the Early Motor Theory was different from the standard two-stage view in a way that had an important effect on the way we thought about speech, and also on the direction of our research. It mattered greatly that we took the object of perception, and the ultimate constituent of phonetic structure, to be an articulatory gesture, not a sound (or its auditory result), for this began a line of thinking that would in time be seen to eliminate the need for the horizontalists' second stage, and so permit us to exorcise the linguistic ghosts—the phonetic names or other cognitive entities--that haunted it. As for the direction of our research, the Early Motor Theory caused us to turn our attention to speech production, and so to initiate the inquiry into that process that has occupied an ever larger and more important place in our enterprise.

Some mildly interesting assumptions that underlay the methods we used

Given the extreme unnaturalness of the stimuli in many of our experiments, and the difficulty subjects reported in hearing them as speech, we had to assume that there would be no important interaction between the degree of approximation to natural speech quality and the way listeners perceive the phonetic information. I therefore note here that this assumption has proved to be correct: no matter how unconvincing our experimental stimuli as examples of speech, those listeners who were nevertheless able to hear them as speech provided results that have held up remarkably well when the experiments were carried out subsequently with stimuli that were closer approximations to the real thing. This was the first indication we had of the theoretically impor-

tant fact that accurate information about the relevant articulator movements, as conveyed, however unnaturally, by the acoustic signal, is sufficient for veridical phonetic perception, provided only that the listener is not too put off by the absence of an overall speech-like quality.

We also had to assume that the validity of our results would be little affected by a practice, followed throughout our search for the cues, of investigating, in any one experiment, only a single phonetic feature in a single phonetic class, and instructing the subjects to limit their responses to the phones in that class. Obviously, this procedure left open the possibility that cues sufficient to distinguish phones along, say, the dimension of place in some particular condition of voicing or manner would not work when voicing or manner was changed, or when the set of stimuli and corresponding response choices was enlarged. In fact, the results obtained in the limited contexts of our experiments were not overturned in later research when, for whatever reason, those limits were relaxed. I take this to be testimony to the independence in perception of the standard feature dimensions.

Finally, given early indications that there were several cues for each phonetic element, and given that, to make our experiments manageable, at least in the early stages, we typically investigated one at a time, we had to assume the absence of strong interactions among the cues. In fact, as we later discovered, there are such interactions—specifically, the trading relations that bespeak a perceptual equivalence among the various cues for the same phone and that are, therefore of some theoretical interest, as I will say later—but these occur only within fairly narrow limits, and they only change the setting of a cue that is optimal for the phone; they do not otherwise affect it. Therefore, working on only one cue at a time did not cause us to be seriously misled.

More context-conditioned variability and the dynamic aspects of the speech signal

The most cursory examination of a spectrogram of running speech reveals nothing so clearly as the almost continuous movement of the formants; even the band-limited noises that characterize the fricatives seem more often than not to be dynamically shaped. Indeed, Potter, Kopp, and Green, in their book, Visible Speech, had remarked these movements, but they considered the effect to proceed from (constant) consonant to (variable) vowel, at least in the case of stop consonant-vowel syllables; and since their interest was primarily in how these transitional influences might help people to 'read' spectrograms, they simply called attention to the direction of the movement, up or down, and did not speculate about the role of these movements in speech perception. Martin Joos, on the other hand, wrote explicitly about the consonant-vowel transitions, as well as their context-conditioned variability, and showed, by cutting out properly selected parts of magnetic-tape recordings, that these transitions conveyed information about the consonants. He also made the important observation that there was, therefore, no direct correspondence between the segmentation of the acoustic signal and the segmentation of the phonetic structure. But Joos could not vary the transition for experimental purposes, so his conclusions could not be further refined. We therefore thought it a reasonable next step to make those variations, and so learn more about the role of the transitions in perception, choosing, first, to study place of production among stop and nasal consonants.

Our research to that point had prepared us to deal only with two-formant patterns, and since inspection of spectrograms indicated that transitions of the first formant did not vary with place, we chose to experiment with transitions of the second. To that end, we varied the starting point, hence the direction and extent, of these transitions by starting them at each of a number of frequencies above and below the steady state of the following vowel. In one condition, the first formant had a fixed transition that rose to the steady state from the very lowest frequency (120 Hz); the resulting patterns were intended to represent voiced stops, and it was our judgment that they did that reasonably well. In a second condition, the first formant had a zero transition—that is, it was straight. We hoped that these patterns would sound voiceless, but, in fact, they did not. They were used nevertheless. The stimuli in each of these conditions were presented for identification as [b], [d], [g] to one group of listeners and as [p], [t], [k] to another. The results showed clearly that the transitions do provide important information about the place dimension of the voiced and voiceless stops, and, also, that this information is independent of voicing. Indeed, it mattered little whether the listeners were identifying a particular set of synthetic stops as voiced or voiceless; the same transitions were associated with the same place, and there was, at most, only slightly less variability in the condition with the rising first formant, where the stimuli sounded voiced to us and the subjects were asked to judge them so. In a

third condition, we strove, fairly successfully we thought, for nasal consonants by using a straight first formant, together with what we considered at the time to be an appropriate (fixed) nasal resonance. Here, too, the second-formant transitions provided information about place, and that information was in no way affected by the change in manner, even though we had reversed the patterns so as to make the nasals syllable final, and so to take account of the fact that the velar nasal never appears initially in the syllable in English. As for context effects, they were large and systematic. Thus, the best transition for [d] (or [t] or [n]) fell from a point considerably above the steady state of the vowel with [u], but with [i] it rose from a point below the vowel's steady state, and similar effects were evident with the transitions for other phones. We thought it supportive of a motor theory that the highly variable transitions for the same consonant were produced by a reasonably constant articulatory gesture as it was merged with the gesture appropriate for the following vowel. Equally supportive, in our view, was the fact that mirror-image transitions in syllable-initial and syllable-final positions nevertheless yielded the same consonantal percept, for surely these would sound very different in any well-behaved auditory system. From an articulatory point of view, however, these transitions are seen as the acoustic consequences of the opening and closing phases of the same gesture. As with the bursts, then, perception cued by the transitions accorded better with an articulatory process than with sound.

Despite the demonstration, by us and others, of the extreme context sensitivity of the acoustic cues, some researchers have been concerned for many years to show that there are, nevertheless, invariant acoustic cues, implying, then, that no special theoretical exertions are necessary in order to account for invariant phonetic percepts. My own view of this matter has always been that, whatever the outcome of the seemingly never-ending search for acoustic invariants, the theoretical issue will remain largely untouched; for there is surely no question that the highly context-sensitive transitions *do* supply important information for phonetic perception—they can, indeed, be shown to be quite sufficient in many circumstances—and that incontrovertible fact must be accounted for.

A brief flirtation with binary decisions

In our first attempt to interpret the significance of the burst and transition results, we took seriously, if only for a short time, the possibility that the two kinds of cues collaborated in such a way that two binary decisions resolved all perceptual ambiguity. Our data had shown that the bursts were identified as [t], if they were high in frequency and as [p] or [k], if low; the second-formant transitions evoked [t] or [k], if they were falling, and [p], if rising. So a low burst and a rising transition would be an unambiguous [p]; a high burst and falling transition would be [t]; and a low burst coupled with a falling transition would be [k]. We were, of course, influenced to this conclusion not just by our data but also, if only indirectly, by the then prevailing fashion for binary arrangements. At all events, we made no attempt to link our notion about binary decisions with the Early Motor Theory, perhaps because that would have been hard to do.

Acoustic loci: Rationalizing the transitions and their role in perception

It required only a little further reflection, combined with an examination of the results of our experiments on the second-formant transitions, to see that perception was sensitive to something more than whether the transition was rising or falling. Since those transitions reflect the cavity changes that occur as the articulators move from the consonant position to the vowel, and, since the place of production for each consonant is more or less fixed, we saw that we should expect to find a correspondingly fixed position—or 'locus', as we chose to call it—for its second formant. More careful examination of the results of our experiments suggested that, for each position on the dimension of place, the transition might, indeed, have originated at some particular frequency, and then made its way to the steady state of the vowel, wherever that was. To refine this notion, we carried out a two-step experiment. In the first, we put steady-state second formants at each of a number of frequency levels, from 3600 Hz to 720 Hz, and paired each with one of a number of first formants in the range 720 Hz to 2400 Hz. The first formants had rising transitions designed to evoke the effect of a voiced stop consonant. Careful listening revealed that [d] was heard when the second formant was at 1800 Hz, [b] at 720 Hz, and [g] at 3000 Hz, so we settled on these frequencies as the loci for the places of production of the three stops.

The second step was to prepare two-formant patterns in which the second formant started at each of these loci, and then rose or fell to the steady state of the vowel, wherever that was. With

these patterns, the consonant appropriate to the locus was not evoked clearly. For [d], indeed, starting the transitions at the locus produced, for some steady states, [b] and [g]. To get good consonants, we had in all cases to 'erase' the first half of the transition so as to create a silent interval between the locus and the actual start of the transition. We noted, further, that in the case of [g], this maneuver worked only with second-formant steady states from 3000 Hz to about 1200 Hz, which is approximately where the vowel shifts from spread to rounded; below 1200 Hz, no approximation to [g] could be heard.

The concept of the locus, together with the experimental results that defined it, made simple sense of the transitions. It also tempted me to temper the emphasis on context-conditioned variability by assuming that the perceptual machinery—by which I might have meant the *auditory* machinery—'extrapolated' backward from the start of the transition to the locus, and so arrived at a 'virtual' acoustic invariant. Fortunately, I yielded to this temptation only briefly, and there is, I think, no written record of my lapse. In any case, we began early to take the opposite tack, using the locus data to strengthen our conclusions about the role of context by emphasising the untoward consequences of actually starting the transitions at the locus, and by pointing to the sudden shift in the [g] locus when the vowel crosses the boundary from spread to rounded.

Stop vs. semivowel, or once more into the auditory breach

It was apparent on the basis of articulatory considerations, and also by inspection of spectrograms, that an acoustic correlate of the stop-semivowel distinction was the duration or rate of the appropriate transitions. To find out how this variable actually affected perception, we carried out the obvious experiment. We particularly wanted to know whether it was rate or duration, and also where on the rate or duration continuum the phonetic boundary was. By varying the positions of the vowel formants, and hence the extent of the transition, we were able to separate the two variables and find that duration seemed to be doing all the work. We also found that the boundary was at about 50 msec.

I recall thinking that 50 msec might be critical for some kind of auditory integration. As I have already said, it seemed reasonable to me at the time to suppose that phonetic distinctions had accommodated themselves to the properties of the auditory system as revealed at the level of psychophysical relations, since the (implicit) motor activity that was ultimately perceived was itself *initially* evoked by some kind of first-stage auditory representation. I was therefore naturally attracted to the possibility that transition excursions of less than 50 msec duration were, perhaps, so integrated by the auditory system as to produce a unitary impression like that of a stop consonant, while transitions with excursions longer than that would evoke the impression of gradual change that characterizes the semivowels. I well remember the excitement I felt when, having decided that a 50-msec duration might well be the auditory key, I appreciated how easy it would be to find out if, indeed, it was. The experiment required only that I draw on the Playback a series of rising and falling isolated transitions in which the duration was varied over a wide range. What I expected and hoped to find was that a duration of 50 msec would provide a boundary between perception of a unitary stop-like burst of sound on the one side, and a semivowel kind of glide on the other. So far as I could tell, however, there appeared to be no such boundary at 50 msec and thus no evidence of an auditory basis for the results of our experiment on the distinction between stop and semivowel. I was disapppointed, but not enough to abandon my horizontal attitude about the role of auditory representations in the ontogenetic development of phonetic perception.

Categorical perception: the right prediction from the wrong theories

According to just those aspects of the Early Motor Theory that I now believe to be mistaken, the auditory percept originally evoked by the speech signal was supposed to give way to the sensory consequences of the articulatory gesture, and it was just those consequences that were ultimately perceived. In arriving at this theory, I had, of course, been much influenced by the behaviorist stimulus-response tradition in which I had been reared. It was virtually inevitable, then, that I should take the next step and consider the consequences of two processes—'acquired distinctiveness' and 'acquired similarity'—that were part of the same tradition. The point was simple enough: if two stimuli become connected, through long association, to very different responses, then the feedback from those responses, having become the end-states of the perceptual process, will cause the stimuli to be more discriminable than they had originally been; conversely, if these stimuli become connected to the same response, then, for

the same reason, they will be less discriminable. In fact, there was not then, and is not now, any evidence that such an effect occurs. But that did not trouble me, for I supposed that investigators had not thought to look in the right place, and I could not imagine a better place than speech perception. Neither was I troubled by what seems to me now the patently implausible assumption, basic to the concepts of acquired distinctiveness and acquired similarity, that the normal auditory representation of a suprathreshold acoustic stimulus could be wholly supplanted, or even significantly affected, by the perceptual consequences of some motor response just because the acoustic stimulus and the motor response had become strongly associated. It was for me compelling that listeners had for many years been making different articulatory responses to stimuli that happened to lie on either side of a phonetic boundary, so, by the terms of the Early Motor Theory and the theory of acquired distinctiveness, that difference should have become more discriminable. On the other hand, those listeners had been making the same articulatory response to equally different stimuli that happened to lie within the phonetic class, so, by the same theories, those stimuli should have been rendered less discriminable.

To test the theories, I thought it necessary only to get appropriate measures of acoustic-cue discriminability. As I know now, one can easily get the effect I sought simply by listening to voiced stops, for example, as the second-formant transition is changed in relatively small and equal steps, for what one hears is, first, several consonants that are almost identical [b]'s, then a rather sudden shift to [d], followed by several almost identical [d]'s, and then, again, a sudden shift, this time to [g]. Though we had the means to make this simple and quite convincing test, we did not think to try it. Instead, I put together two wrong theories and produced what my professors had taught me to strive for as one of the highest forms of scientific achievement: a real prediction.

The test of the prediction was initially undertaken by one of our graduate students, Belver Griffith, who, with my enthusiastic approval, elected to do the critical experiment on steady-state vowels. We know now that the effect we were looking for does not occur to any significant extent with such vowels, so it was fortunate that Griffith, by nature very fussy about the materials of his experiments, was unsuccessful in producing vowels he was willing to use. The happy consequence was that he, together with the rest of us, decided to move ahead, in parallel, with stop consonants.

It was not our purpose to obtain discrimination thresholds, but only to measure discriminabililty of a constant physical difference at various points on the continuum of second-formant transitions. To that end, we synthesized a series of 14 syllables in which the starting point of the second formant was varied in steps of 120 Hz, from a point 840 Hz below the steady state of the following vowel to a point 720 Hz above it. We then paired stimuli that were one, two, and three steps apart on the continuum, and for each such pair measured discriminability by the ABX method (A and B were members of the pair, X was one or the other, and the subject's task was to match X with A or B). The result was that there were peaks in the discrimination functions at positions on the continuum that corresponded to the phonetic boundaries as earlier determined by the way the subjects had identified the stimuli as [b], [d], or [g] when they were presented singly and in random order. This is to say that, other things equal, discimination was better between stimuli to which the subjects habitually gave different articulatory responses than it was between stimuli to which the responses were the same. Thus, the prediction was apparently confirmed by this instance of what we chose to call 'categorical perception'.

In the published paper, we included a method, worked out by Katherine Harris, for computing from the absolute identification functions what the discrimination function would have been if, indeed, perception had been perfectly categorical—that is, if listeners had been able to perceive differences only if they had assigned the stimuli to different phonetic categories. Applying this calculation to our results, we found that, in this experiment at least, perception was rather strongly categorical, but not perfectly so.

To this point in our research, psychologists, including even those interested in language, had paid us little attention. Requests for reprints, and such other tokens of interest as we had received, had come mostly from communication engineers and phoneticians, and there were few references to our work in the already considerable literature of psycholinguistics, a field that had been established, seemingly by fiat, by a committee of psychologists and linguists who had met for a summer work session at Cornell. The result of their deliberations was a briefly famous monograph in which they defined the new

discipline, constructed its theoretical framework, and posed the questions that remained to be answered. That done, they officially launched the field at a symposium held during the next national convention of the American Psychological Association. At the end of the symposium, I asked a question that provoked one of the founding fathers to inform me, icily, that speech had nothing to do with psycholinguistics. He did not say why, but then, as one of the inventors of the discipline, he was entitled to speak ex cathedra. It is, however, easy to appreciate that in looking at speech horizontally, as he and the other members of the committee surely did, one sees nothing that is linguistically interesting, only a set of unexceptional noises and equally unexceptional auditory percepts that just happen to convey the more invitingly abstract structures where anyone who would think deeply about language ought properly to put his attention.

The categorical perception paper seemed, however, to touch a psycholinguistic nerve. Perhaps this was because, if taken seriously, it showed that the phonetic units were categorical, not only in their communicative role, but also as immediately perceived; and this nice fit of perceptual form to linguistic function must have seemed at odds with the conventional horizontal assumption that the auditory percepts assume linguistic significance only after a cognitive translation, not before. Our results could be taken to imply that no such translation was necessary, and that there might, therefore, be something psycholinguistically interesting and important about the precognitive—that is, purely perceptual—processes by which listeners apprehend phonetic structures.

Most psychologists seemed unwilling to accept that implication, though not all for the same reason. Some argued that categorical perception, as we had found it, was of no consequence because it was merely an artifact of our method. This criticism boiled down to an assertion, perfectly consistent with the standard horizontal view, that the memory load imposed by the ABX procedure made it impossible for the subject to compare the initial auditory representations, forcing him to rely, instead, on the categorical phonetic names he assigned to the rapidly fading auditory echoes as they were removed from the everchanging sensory world and elevated, for safer keeping, into short-term memory. It is, of course, true that reducing the time interval between the stimuli to be compared does raise the troughs that appear in the within-category parts of the discrimination funcntion, and thus reduces the approximation to categorical perception. But the Early Motor Theory did not require that the articulatory responses within a phonetic category be identical, so it did not predict that perception had to be perfectly categorical. (The degree to which the articulatory responses within a category are similar presumably varies according to the category and the speaker; it is, therefore, a matter for empirical determination.) Neither did the theory in any way preclude the possibility that perceptual responses would be easier to discriminate when fresh than when stale. In any case, it surely was relevant that the peaks one finds with various measures of discrimination merely confirm the quantal shifts a listener perceives as the stimuli are moved along the physical continuum so rapidly as to make the memory load negligible.

The other criticism, which seemed almost opposite to the one just considered, was that categorical perception is not relevant to psycholinguistics because it is so common, and, more particularly, because those boundaries that the discrimination peaks mark are simply properties of the general auditory system, hence not to be taken as support for the view that speech perception is interesting, except, perhaps, within the domain of auditory psychophysics. As for the criticism that categorical perception is common, it seemed to have been based on the misapprehension that we had claimed categorical perception to be unique to speech, but in fact we had not, having merely observed (correctly) that, given stimuli that lie on some definable physical continuum, observers commonly discriminate many more than they can identify absolutely. Our claim about phonetic perception was only that there is a significant, if nevertheless incomplete, tendency for that commonly observed disparity to be reduced. On the other hand, the claim by our critics that the boundaries are generally auditory, not specifically phonetic, was important and deserved to be taken seriously. It has led to many experiments on perception of nonspeech control stimuli and on perception of speech by nonhuman animals, leaving us and the other interested parties with an issue that is still vexed. In fact, I think the weight of evidence, taken together with arguments of plausibility, overwhelmingly favors the conclusion that the boundaries are specific to the phonetic system, but I reserve the justification for that conclusion to the last section of the paper.

There was yet another seemingly widespread misapprehension about categorical perception, which was that it had served us as the primary

basis for the Early Motor Theory. In fact, we (or, at least, I) have long believed that the facts about this phenomenon are consistent with the theory, but they do not by any means provide its most important support; indeed, they were not available until at least five years after we had been persuaded to the theory, as I earlier indicated, by the very first results of our search for the acoustic cues.

Finally, in the matter of categorical perception, I will guess that consonant perception is likely, when properly tested, to prove more nearly categorical than experiments have so far shown it to be. The problem with those experiments is that they have used acoustic synthesis, so it has been prohibitively difficult in any single experiment to make proper variations in more than one of the many aspects of the signal that are perceptually relevant. But when only one cue is varied, as in the experiments so far done, then, as it is changed from the form appropriate for one phone to the form appropriate for the next, it leaves all the other relevant information behind, as it were, creating a situation in which the listener is discriminating, not just the phonetic structure, but also increasingly unnatural departures from it. I suspect that, with proper articulatory synthesis, when the acoustic signal will change in all relevant aspects—at least for the cases that are produced by articulations that can be said to vary continuously—the discrimination functions will come much closer to being perfectly categorical.

The concept of 'cue' as a theoretically relevant entity

At the very least, 'cue' is a term of convenience, useful for the purpose of referring to any piece of signal that has been found by experiment to have an effect on perception. We have used the word in that sense, and continue to do so. But there was a time when cue had, at least in my mind, a more exalted status. I supposed that there was, for each phone, some specifiable number of particulate cues that combined according to a scientifically important principle to evoke the correct response. It was this understanding of cues that was implicit in the 'binary' account of their effects that I referred to earlier. The same understanding was more explicit in a dissertation on cue combination that I had urged on Howard Hoffman. Finally, and perhaps most egregiously, it became the centerpiece of our interpretation of an experiment on the effects of third-formant transitions on perception of place among the stops. Having found there that, in enhancing the perception of one stop, any particular transition does not do so equally at the expense of the other two, we concluded that a cue not only tells a listener what a speech sound is, but also which of the remaining possibilities it is not. It was almost as if we were supposing that the third-formant transition had been designed, by nature or by the speaker, just to resolve an ambiguity that the more important second-formant transition had overlooked. At all events, we went on to speculate that the response alternatives exist in a multidimensional phonetic space, and, though we were not perfectly explicit about this in the published paper, that a cue has a magnitude and a direction, just like a vector, with the result that the final position of the percept in the phonetic space is determined by the sum of the vectors. Such a conception is, of course, at odds with all the data now available that indicate how exquisitely sensitive the listener is to *all* the acoustic consequences of phonetically significant gestures, for what those data mean is that any definition of an acoustic cue is always to some extent arbitrary. Surely, it makes little sense to wonder about the rules by which arbitrarily defined cues combine to produce a perceptual result.

The voicing distinction; an exercise in not seeing that which is most visible

We discovered very early how to produce stops that were convincingly voiced, but we had been frustrated for five years or more while seeking the key to synthesis of their voiceless counterparts. In our quest, we had examined spectrograms, sought advice from colleagues in other laboratories, and, by trial and error on the Playback, tried every trick we could think of. We varied the strength of the burst relative to the vocalic section of the syllable, drew every conceivable kind of first-formant transition, and substituted various intensities of noise for the harmonics through varying lengths of the initial parts of the formant transitions.

In fact, there was no noise source in the Pattern Playback, only harmonics of a 120 Hz fundamental, but we had been able in research on the fricatives to make do by placing small dots of paint where noise was supposed to be. In isolation, patches of such dots sounded like a twittering of birds, but in syllabic context they produced fricatives so acceptable that, when we used them in experiments, we got results virtually identical to those obtained later when, with a new synthesizer called Voback, we were able to deploy

real noise. Before Voback was available, however, we had, in our attempts at voiceless stops, to rely on the trick that had worked for the fricatives. When it did not help, we concluded that, unlike the fricatives, the voiceless stops needed true noise and, accordingly, that our inability to synthesize them was to be attributed to the noise-producing limitations of the Playback.

That we were wrong to blame the Playback became apparent one day as a consequence of a discovery by Andre Malecot, one of Pierre's graduate students. While working to synthesize the syllable-final releases of stops, he omitted the first formant of the short-duration syllable that constituted the release. I believe that he did this inadvertently. But whether by inadvertence or by design, he produced a dramatic effect: we all heard a stop that was quite clearly voiceless. Encouraged by this finding, we adapted it to stops in syllable initial position, and carried out several related experiments. In each, we varied one potential cue for the voicing contrast for all three stops, paired with each of the vowels [i], [ae], and [u]. Our principal finding was that, with all else equal, simply delaying the onset of the first formant relative to the second and third was sufficient to cause naive listeners to shift their responses smartly between voiced and voiceless. Then, recognizing that in so delaying the onset of the first formant we were, at the same time, starting it at a higher frequency, we reconfirmed the observation we had made in our earlier experiments, which was that starting the first formant at a very low frequency was important in creating the impresssion of a voiced stop, but that starting it higher, at the steady state of that formant, did not, by itself, make much of a contribution to voicelessness. On the other hand, delaying the onset of the first formant without at the same time raising its starting frequency did prove to be a very potent cue. Indeed, it appeared from the responses of our listeners to be about as potent as the original combination of delay and raised starting point. (For the purpose of this experiment, we varied the delay alone by contriving a synthetic approximation to the vowel [o] in which the first formant was placed as low on the pattern as it could go; it was, then, just this straight formant that was delayed.) Next, we took advantage of the newly available synthesizer, Voback, which, as I earlier said, had a proper noise source, to experiment with the effect of noise in place of harmonics during the transitions. What we found was that substituting noise in all three formants was, by itself, ineffective, but that substituting it for harmonics in the second and third formants for the duration of the delay in first-formant onset did somewhat strengthen the impression of voicelessness. In connection with this last conclusion, we noted that when, in an attempt to produce an initial [h], which is, of course, the essence of aspiration, we replaced the harmonics of all the formants with noise for the first 50 or 100 msec, we did not get [h], but rather the impression of a vowel that changed from whispered to fully voiced; to get [h], we had to omit the first formant. To explain all this, we advanced a suggestion, made to us by Gunnar Fant, that the vocal cords were open during the period of aspiration, and that it was this circumstance that reduced the intensity of the first formant, thus effectively delaying its onset. We emphasized, then, that all the acoustically diverse cues were consequences of the same articulatory event, and therefore led, in accordance with the Motor Theory, to a percept that was perfectly coherent.

This early work on the voicing distinction was subsequently refined and considerably extended by Arthur Abramson and Leigh Lisker. In particular, it was they who established how the acoustic boundaries for the voicing distinction vary with different languages, and thus provided the basis for the great volume of later research by other investigators who exploited these differences in pursuit of their interests in the ontogenesis of speech. And it was Abramson and Lisker who accurately characterized the relevant variable as voice-onset-time (VOT), defined as the duration of the interval between the consonant opening (in the oral part of the tract) and the onset of voicing at the larynx. Unfortunately, some of the researchers who later used the voicing distinction for their own purposes ignored the fact that the VOT variable is articulatory, not acoustic, and therefore failed to take into account in their theoretical interpretations that its acoustic manifestations are complexly heterogeneous.

As for our initial discovery of the acoustic cues for the voicing distinction, I note an irony in the long search that preceded it, for once one knows where to look in the spectrogram, the delay in the first formant onset can be measured more easily, and with greater precision, than almost any of the other consonant cues we had found. Consider, for example, how important to perception is the frequency at which a formant transition starts, and then how hard it is to specify that frequency precisely from an inspection of its appearance on a spectrogram. Yet our tireless examination of

spectrograms had, in the case of the voicing distinction, availed us nothing; we simply had not seen what we now know to be so plainly there.

Synthesis by rule and a reading machine that speaks

In this very personal chronicle of our early research, I have chosen to write of just those experiments that best illustrate certain underlying assumptions I now find interesting. I have said nothing about the many other experiments that were carried out during roughly the same period of time. I would now partly repair that omission by recognizing the existence of those others, and by emphasizing that they provided a collection of data sufficient as a basis for synthesizing speech from a phonetic transcription, without the need to copy from a spectrogram. Unfortunately, all the relevant data had been brought together only in Pierre's head. Relying only on the experience he had gained from participation in our published research, and also from the countless unpublished experiments he had carried out in his unflagging effort to refine and extend, Pierre could 'paint' speech to order, as it were. But the knowledge that Pierre had in his head was, by its nature, not so public as science demands.

We therefore recommended to Frances Ingemann. when she came to spend some time at the Laboratories, that she write a set of rules for synthesis, making everything so explicit that someone totally innocent of knowledge about acoustic phonetics could, simply by following the rules, draw a spectrogram that would produce any desired phonetic structure. Accepting this challenge, she decided to rely entirely on the papers we had published in journals or in lab reports; she did not use the synthesizer to test and improve as she went along, nor did she attempt to formalize what Pierre and other members of the staff might know but had never written down. She nevertheless succeeded very well, I think, in producing what must count as the first rules for synthesis. I don't recall that we ever formally assessed the intelligibility of the speech produced by these rules, but I know that we found it reasonably intelligible. At all events, an outline of the rule system was published in 1959 under the title, 'Minimal Rules for Synthesis'. The word 'minimal' was appropriate because the rules were written at the level of features, the presumed 'atoms' of phonetic structure, not at the level of the more numerous segments or 'molecules'. I should note, too, that we took explicit notice in the paper of our belief that the rules for synthesis had better be written in articulatory terms, for then *all* the relevant acoustic information would be provided to the listener. There was, however, no alternative to the acoustically based rules we offered, because there was not enough known about articulation, but also because there was, in any case, no satisfactory articulatory synthesizer. Articulatory synthesis would therefore have to wait.

Meanwhile, we had, by 1966, a computing facility, and had constructed a computer-controlled, terminal-analog formant synthesizer. It was then that Ignatius Mattingly joined the staff and undertook to program the computer to produce speech by rule. For this purpose, he drew on the work he had done previously with Holmes and Shearme in England, and also on all that had been learned about the cues and rules for synthesis at the Laboratories. By 1968 the job was done. Accepting an input of a phonetic string, the system would speak. The intelligibility of the speech was tested on several occasions and in several different ways. Thus, it was tested informally by having blind veterans listen, for example, to rather long passages from Dickens and Steinbeck. It was evident that they understood the speech, even at rates of 225 words per minute, but we had no measure of exactly how hard they found it to do so, and they did complain, not without reason we thought, of what they called the machine's 'accent'. In more formal tests, the rule-generated synthetic speech came off quite well by comparison with 'real' speech, but, not unexpectedly, there was evidence of a price exacted by the extra cognitive effort that was required to overcome its evident imperfections, a price that had to be paid, presumably, by the processes of comprehension.

At that point, we had in hand a principal component of a reading machine that would convert text to speech, and thus avoid all the problems we had encountered in our earlier work with nonspeech substitutes. What was needed, in addition, was an optical character reader to convert the letters into machine-readable form, and also, of course, some way of translating spelled English into a phonetic transcription appropriate to the synthesizer. Given our history, it was inevitable that we should have been impatient to acquire these other components and see (or, more properly, hear) what a fully automatic system could do. So, Frank, Patrick Nye, and others cobbled together just such a system, using an optical character reader we bought with money given us for the purpose by the Seeing Eye Foundation, a phonetic dictionary

made available by the Speech Communications Research Laboratory of Santa Barbara, and, of course, our own computer-controlled synthesizer. Tests revealed that the speech produced in this fully automatic way was almost as good as that for which the phonetic transcription had been hand-edited. But we were concerned about the evidence we had earlier collected concerning the probable consequences for ease of comprehension that arose out of the shortcomings of the speech. We therefore put together a plan to evaluate the machine with blind college students who would use it to read their assignments; having found its weaknesses, we would then try to correct them. I assumed that various federal agencies would compete to see which one could persuade us to accept their support for this undertaking. We could, after all, show that a reading machine for the blind was not pie-in-the-sky, but a do-able thing that stood in need of just the kinds of improvement that further research would surely bring. Yet, though we tried very hard with several agencies, and for several years, we failed utterly to get support, and were forced finally to abandon our plans. Still, we had the satisfaction of having proved to ourselves that a reading machine for the blind was close to being a reality. The basic research was largely complete; what remained was just the need for proper development.

ON BECOMING VERTICAL, OR HOW I RIGHTED MYSELF

To this point, my concern has been to describe the various forms of the horizontal view that my colleagues and I held during our work on non-speech reading machines and in the early stages of the research on speech to which it led. Now I mean to offer a more detailed account of the important differences between that view and the vertical view I now hold. In so doing, I draw freely, and without specific attribution, on a number of theoretically oriented papers that were written in close collaboration with various of my colleagues. Among the most relevant of these are several reviews by Ignatius Mattingly and me in which we hammered out the vertical view as I (we) see it now.

All these theoretically oriented papers deal, at least implicitly, with questions about speech to which the horizontal and vertical views give different, sometimes diametrically opposed, answers. Such questions serve well, therefore, to define the two positions, and to explain how I came to abandon the one for the other; for those reasons, I will organize this section of the paper around them.

The issue that unites the questions pertains to the place of speech in the biological scheme of things. That I should have come to regard that issue as central is odd, given the habits of mind I had brought to the research, for, as I earlier implied, my education in psychology had been unremittingly abiological. I had, to be sure, studied a little physiology, narrowly conceived, and it cannot have escaped my notice that in the physiological domain things were not of a piece, having been formed, rather, into distinct systems for correspondingly distinct functions. At the level of behavior, however, I saw only an overarching sameness, a reflection of my attachment to principles so general as to apply equally to a process as natural as learning to speak and as arbitrary as memorizing a list of nonsense syllables.

I think I was moved first, and most generally, to a different approach by scientists who work, not on speech, but on other forms of species-typical behavior. Thus, it was largely under the influence of people like Peter Marler, Nobuo Suga, Mark Konishi, and Fernando Nottebohm that I came to see myself as an ethologist, very much like them, and to appreciate that I would be well advised to begin to think like one. They helped me to understand that speech is to the human being as echolocation is to the bat or song is to the bird—to see, that is, that all these behaviors depend on biologically coherent faculties that were specifically adapted in evolution to function appropriately in connection with events that are of particular ecological significance to the animal. To the horizontalist that I once was, this was heresy; but to the verticalist I was in process of becoming, it was the beginning of wisdom.

Meanwhile, back at the Laboratories there were biological stirrings on the part of Michael Studdert-Kennedy, who is nevertheless not a committed verticalist, and Ignatius Mattingly, who is. Michael has been a constructive critic in regard to virtually every biologically relevant notion I have dared to entertain. As for the biological slant of the vertical view (including the Revised Motor Theory), that is as much Ignatius's contribution as mine. Indeed, the view itself is the result of a joint effort, though, of course, he bears no responsibility for what I say about it here.

Among the influences of a somewhat different sort, there was the growing realization that my early horizontal view did not sit comfortably even with the results of the early research it was

designed to explain. That will have been seen in what I have already said about my attempts to account for those results, and, especially, about the patch on the horizontal view that I have here called the Early Motor Theory. Heavily loaded as it was with untested and wholly implausible assumptions—for example, that auditory percepts could, as a result of learning, be replaced by sensations of movement—it had begun to fall of its own weight.

Contributing further to the collapse of the Early Motor Theory was the research, pioneered by Peter Eimas and his associates, in which it was found that prelinguistic infants had a far greater capacity for phonetic perception than a literal reading of the theory would allow.

My faith was further weakened by the work of Katherine Harris and Peter MacNeilage, who, as the first of the Laboratories' staff to work on speech production, were busily finding a great deal of context-conditioned variability in the peripheral articulatory movements (as reflected in electromyographic measures), and thus disproving one of the assumptions of the Early Motor Theory, which was that the articulatory invariant was in the final-common-path commands to the muscles.

At the same time, Michael Turvey was pointing the way to an appropriate revision by showing how, given context-conditioned variability at the level of movement, it is nevertheless possible, indeed necessary, to find invariance in the more remote motor entities that Michael called 'coordinative structures'. In any case, Michael and Carol Fowler were strongly encouraging me to persevere in the aspect of the Early Motor Theory that took gestures to be the objects of speech perception, while simultaneously heaping scorn on the idea that perception was a second-order translation of a sensory representation, as the horizontal version of the theory required. I began, therefore, to take more seriously the possibility that there is no mediating auditory percept, only the immediately perceived gestures as provided by a system—the phonetic module—that is specialized for the ecologically important function of representing them.

Not that Michael and Carol or, indeed, any of the other 'ecological' psychologists in the Laboratories, are verticalists. They most certainly are not, because they do not accept (yet) that there is a distinct phonetic mode, preferring, rather, to take speech perception as simply one instance of the way all perception is tuned to perceive the distal objects; in the case of speech, these just happen to be the articulatory gestures of the speaker. Thus, I have been in the happy position of taking advantage of the best of what my ecological friends have had to offer, while freely rejecting the rest, and, as an important bonus, being stimulated by our continuing disagreements to correct weaknesses and repair omissions in my own view.

It was also relevant to the development of my thinking that Isabelle Liberman, Donald Shankweiler, and Ignatius Mattingly—followed later by such younger colleagues as Benita Blachman, Susan Brady, Anne Fowler, Hyla Rubin, and Virginia Mann—had begun to see in our research how to account for the fact that speech is so much more natural (hence easier) than reading and writing, and thus to be explicit about what is required of the would-be reader/writer that mastery of speech will not have taught him. As I will say later, their insights and the results of their empirical work illluminated aspects of the vertical view that I would otherwise not have seen.

I was affected, too, by the results of experiments on duplex perception, trading relations, and integration of cues, experiments that went beyond those, referred to earlier, that merely isolated the cues and looked for discontinuities in the discrimination functions. These later experiments, done (variously) in close collaboration with Virginia Mann, Bruno Repp, Douglas Whalen, Hollis Fitch, Brad Rakerd, Joanne Miller. Michael Dorman, and Lawrence Raphael (few of whom are admitted verticalists) provided data that spoke more clearly than the earlier findings to some of the shortcomings of the horizontal position, and therefore inclined me ever more strongly to the vertical alternative.

Finally, I should acknowledge the profound effect of Fodor's provocative monograph, 'The Modularity of Mind', which, in the early stages of my conversion, enlightened and stimulated me by its arguments in favor of the most general aspects of the vertical view.

That I should finally have asked the following questions, and answered them as I do, reflects the influences I have just described, and fairly represents the theoretical position to which they moved me.

In the development of phonological communication, what evolved?

Defined as the production and perception of consonants and vowels, speech, as well as the phonological communication it underlies, is plainly a species-typical product of biological

evolution. All neurologically normal human beings communicate phonologically; no other creatures do. The biologically important function of phonologic communication derives from the way it exploits the combinatorial principle to generate vocabularies that are large and open, in contrast to the vocabularies of nonhuman, nonphonologic systems, which are small and closed. Thus, phonological processes are unique to language and to the human beings who command them. It follows that anyone who would understand how speech works must answer the question: what evolved? Not when, or why, or how, or by what progression from earlier-appearing stages. The first question is simply: what?

The answer given by the horizontal view is clear: at the level of action and perception, nothing evolved; language simply appropriated for its own purposes the plain-vanilla ways of acting and perceiving that had developed independently of any linguistic function. Thus, those horizontalists who put their attention rather narrowly on the perceptual side of the process argue that the categories of phonetic perception simply reflect the way speech articulation has accommodated itself to the production of sounds that conform to the properties of the auditory system, a claim that I will evaluate in some detail later. A recent and broader, but still horizontal, take on the same issue distributes the emphasis more evenly between production and perception, arguing that phonetic gestures were selected by language on the basis of constraints that were generally motor, as well as generally auditory, However, the important point in this, as in the narrower view, is that the constraints are independent of a phonetic function, hence in no way specific to speech. Put forth as an explicit challenge to the vertical assumption, the broader view has it that there is no reason to assume a special mode for the production and perception of speech, if, with the proper horizontal orientation, one can see that the units of speech are optimized with respect to motor and perceptual constraints that are biologically general.

But the question is not whether language somehow developed out of the biology that was already there; surely, it could hardly have done otherwise. The question, to put it yet again, asks, rather, what did that development produce as the basis for a unique mode of communication? When the horizontalists say that the development of this mode was accomplished merely by a process of selection from among the possibilities offered by general faculties that are independent of language, they are giving an account that applies as well to the development of, say, a cursive writing system. Was not the selection of the cursive gestures similarly determined by motor and perceptual constraints that are independent of language? Yet, what that selection produced were not the biologically primary units of speech, but only a set of optical artifacts that had then to be connected to speech in a wholly arbitrary way. Of course, this is merely to say the obvious about the relation between speech and a writing system, which is that the evolution of the one was biological, the other, not. That is surely a critical difference, but one that the horizontal view must have difficulty comprehending.

If pressed further to answer the question about the product of evolution, the horizontalists would presumably have to say that, while nothing evolved at the level of perception and action, there must have been relevant developments at a higher cognitive level. Thus, it would have been evolution that produced the phonetic entities of a cognitive type to which the nonphonetic acts and percepts of speech must, on the horizontal view, be associated. Being neither acts nor percepts, these cognitive entities—or ideas, as they might be—would presumably be acceptable within the horizontal framework as genetically determined adaptations for language, hence special in a way that speech is not allowed to be. In itself, this seems an unparsimonious, not to say biologically implausible, assumption. And it can be seen to be the more unparsimonious and implausible once the horizontalist tries to explain how the phonetically neutral acts and percepts got connected to the specialized cognitive entities in the first place. In the case of a script, to wring one more point out of that tired example, the obviously nonphonetic motor and visual representations of the writer and reader were connected to language by agreement among the interested parties. Can we seriously propose a similar account for speech?

If the horizontalists should reject the notion that phonetic ideas were the evolutionary bases for speech, there remains to them the most thoroughly horizontal view of all, which is that what evolved was a large brain. In that case, they might suppose either that phonological communication was an inevitable by-product of the cognitive power that such a brain provides, which seems unlikely, or that phonological communication was an invention, created by large-brained people who were smart enough to have appreciated the immense advantages for communication of the combinatorial principle, which seems absurd.

The vertical view is different on all counts. What evolved, on this view, was the phonetic module, a distinct system that uses its own kind of signal processing and its own primitives to form a specifically phonetic way of acting and perceiving. It is, then, this module that makes possible the phonological mode of communication.

The primitives of the module are gestures of the articulatory organs. These are the ultimate constituents of language, the units that must be exchanged between speaker and listener if linguistic communication is to occur. Standing apart as a class from the nonphonetic activities of the same organs—for example, chewing, swallowing, moving food around in the mouth, and licking the lips—these gestures serve a phonetic function and no other. Hence they are marked by their very nature as exclusively phonetic in character; there is no need to make them so by connecting them to linguistic entities at the cognitive level. As part of the larger specialization for language, they are, moreover, uniquely appropriate to other linguistic processes. Thus, the syntactic component is adapted to operate on the specifically phonetic representations of the gestures, not on representations of an auditory kind. Indeed, it is precisely this harmony among the several components of the language specialization that makes the epithet 'vertical' particularly apposite for the view I am here promoting.

Of course, the gestures constitute only the phonetic structures that the perceptual process extracts from the speech signal. Such aspects of the percept as, for example, those that contribute to the perceived qualilty of the speaker's voice are not part of the phonetic system. Indeed, these are presumably auditory in the ordinary sense, except as they may figure in speaker identification, for which there may be a separate specialization.

It is not only the gestures themselves that are specifically phonetic, but also, presumably, their control and coordination. Surely, there is in speech production, as in all kinds of action, the need to cope with the many-to-one relations between means and ends, and also to reduce degrees of freedom to manageable proportions. In these respects, then, the management of speech and nonspeech movements should be subject to the same principles. But there is, in addition, something that seems specific to speech: the grossly overlapped and smoothly merged movements at the periphery are controlled by, and must preserve information about, relatively long strings of the invariant, categorical units that speech cares about but other motor systems do not. And, certainly, it is relevant to the claim about a specialized mode of production that speech, in the very narrowest sense, is species-specific: given every incentive and opportunity to learn, chimpanzees are nevertheless unable to manage the production of simple CVC syllables. (The fact that the dimensions of their vocal tracts presumably do not allow a full repertory of vowels should not, in itself, preclude the articulation of syllables with whatever vowels their anatomy permits.)

As for the evolution of the phonetic gestures, I should think an important selection factor was not so much the ease with which they could be articulated, or the auditory salience of the resulting sound, but rather how well they lent themselves to being coarticulated. For it is coarticulation that, as I will have occasion to say later, makes phonological communication possible.

But it is also this very coarticulation that, as we saw earlier, frustrates the attempt to find the phonetic invariant in the acoustic signal or in the peripheral movements of the articulators. Still, such motor invariants must exist, not just for the aspect of the Motor Theory that explains how phonetic segments are perceived, but for just any theory that presumes to explain how they are produced; after all, speech does transmit strings of invariant phonological structures, so these invariants must be represented in some way and at some place in the production process. But how are they to be characterized, and where are they to be found? Having accepted the evidence that they are not in the peripheral movements, as the Early Motor Theory assumed, Mattingly and I proposed in the Revised Motor Theory that attention be paid instead to the configurations of the vocal tract as they change over time and are compared with other configurations produced by the same gesture in different contexts. As for the invariant causes of these configurations, they are presumably to be found in the more remote motor entities—something like Turvey's coordinative structures—that control the various articulator movements so as to accomplish the appropriate vocal-tract configurations. It is, I now think, structures of this kind that represent the phonetic primitives, providing the physiological basis for the phonetic intentions of the speaker and the phonetic percepts of the listener. Unfortunately for the Motor Theory, we do not yet know the exact characteristics of these motor invariants, nor can we adequately describe the processes by which they control the movements of the articulators. My colleagues, including especially Cathe

Browman, Louis Goldstein, Elliot Saltzman, and Philip Rubin, are currently in search of those invariants and processes, and I am confident that they will, in time, succeed in finding them. Meanwhile, I will, for all the reasons set forth in this paper, remain confident that motor invariants do exist, and that they are the ultimate constituents of speech, as produced and as perceived.

According to the Revised Motor Theory, then, there is a phonetic module, part of the larger specialization for language, that is biologically adapted for two complementary processes: one controls the overlapping and merging of the gestures that constitute the phonetic primitives; the other processes the resulting acoustic signal so as to recover, in perception, those same primitives. On this view, one sees a distinctly linguistic way of doing things down among the nuts and bolts of action and perception, for it is there, not in the remote recesses of the cognitive machinery, that the specifically linguistic constituents make their first appearance. Thus, the Revised Motor Theory is very different from its early ancestor; the two remain as one only in supposing that the object of perception is the invariant gesture, not the context-sensitive sound.

How is the requirement for parity met?

In all communication, whether linguistic or not, sender and receiver must be bound by a common understanding about what counts: what counts for the sender must count for the receiver, else communication does not occur. In the case of speech, speaker and listener must perceive, or otherwise know, that, out of all possible signals, only a particular few have linguistic significance. Moreover, the processes of production and perception must somehow be linked; their representations must, at some point, be the same. Though basic, this requirement tends to pass unnoticed by those who look at speech horizontally, and, especially, by those whose preoccupation with perception leaves production out of account. However, vertical Motor Theorists like Ignatius Mattingly and me are bound to think the requirement important, so we have given it a name—'parity'—and asked how, in the case of speech communication, it was established and how maintained.

Horizontalists must, I think, find the question very hard. For if, as their view would have it, the acts of the speaker are generally motor and the percepts of the listener generally auditory, then act and percept have in common only that neither has anything to do with language. The horizontalist is therefore required to assume that these representations are linked to language and to each other only insofar as speaker and listener have somehow selected them for linguistic use from the indefinitely large set of similarly nonphonetic alternatives, and then connected them at a cognitive level to the same phonetic name or other linguistic entity. Altogether, a roundabout way for a natural mode of communication to work.

For the verticalists, on the other hand, the question is easy. On their view, it was specifically phonetic gestures that evolved, together with the specialized processes for producing and perceiving them, and it is just these gestures that provide the common currency with which speaker and listener conduct their linguistic business. Parity is thus guaranteed, having been built by evolution into the very bones of the system; there is no need to arrive at agreements about which signals are relevant and how they are to be connected to units of the language.

How is speech related to other natural modes of communication?

I noted earlier that human beings communicate phonologically but other creatures do not, and, further, that this difference is important, because it determines whether the inventory of 'words' is open or closed. Now, in the interest of parsimony, I ask whether either view of speech allows that there is, nevertheless, something common to two modes of communication that are equally natural.

On the horizontal view, the two modes must be seen as different in every important respect. Nonhuman animals, the horizontalists would presumably agree, communicate as they do because of their underlying specializations for producing and perceiving the appropriate signals. I doubt that anyone would seriously claim that these require to be translated before they can take on communicative significance. For the human, however, the horizontal position, as we have seen, is that the specialization, if any, is not at the level of the signal, but only at some cognitive remove. I find it hard to imagine what might have been gained for human beings by this evolutionary leap to an exclusively cognitive representation of the communicative elements, except, perhaps, the smug satisfaction they might take in believing that they communicate phonologically, and the nonhuman animals do not, because they have an intellectual power the other creatures lack, and that even in the most basic aspects of communication they can count themselves broad generalists, while the others must be seen as narrow specialists.

On the vertical view, human and nonhuman communication alike depend on a specialization at the level of the signal. Of course, these specializations differ one from another, as do the vehicles—acoustic, optical, chemical, or electrical—that they use. And, surely, the phonetic specialization differs from all the others in a way that is, as we know, critical to the openness or generativity of language. Still, the vertical view permits us to see that phonetic communication is not completely off the biological scale, since it is, like the other natural forms of communication, a specialization all the way down to its roots.

What are the (special) requirements of phonological communication, and how are they met?

If phonology is to use the combinatorial principle, and so serve its critically important function of building a large and open vocabulary out of a small number of elements, then it must meet at least two requirements. The more obvious is that the phonological segments be commutable, which is to say discrete, invariant, and categorical. The other requirement, which is only slightly less obvious, concerns rate. For if all utterances are to be formed by stringing together an exiguous set of commutable elements, then, inevitably, the strings must run to great lengths. There is, therefore, a high premium on rapid communication of the elements, not only in the interest of getting the job done in good time, but also in order to make things reasonably easy, or even feasible, for those other processes that have got to organize the phonetic segments into words and sentences.

Consider how these requirements would be met if, as the horizontal view would have it, the elements were sounds and the auditory percepts they evoke. If it were these that had to be commutable, then surely it would have been possible to make them so, but only at the expense of rate. For sounds and the corresponding auditory percepts to be discrete, invariant, and categorical would require that the segmentation be apparent at the surface of the signal and in the most peripheral aspects of the articulation. How else, on the horizontal view, could commutability be achieved, except as each discrete sound and associated auditory percept were produced by a correspondingly discrete articulatory maneuver? Of course, the sounds and the percepts might be joined, as are the segments of cursive writing, and that might speed things up a bit, but, exactly as in cursive writing, the segmentation would nevertheless have to be patent. The consequence would be that, to say a monosyllabic word like 'bag', the speaker would have to articulate the segments discretely, and that would produce, not the monosyllable 'bag', but the trisyllable [bɔ] [æ] [gɔ]. To articulate the syllable that way is not to speak, but to spell, and spelling would be an impossibly slow and tedious way to communicate language.

One might imagine that if production had been the only problem in the matter of rate, nature might have solved it by abandoning the vocal tract, providing her human creatures, instead, with acoustic devices specifically adapted to producing rapid-fire sequences of sound. That would have taken care of the production problem, while, at the same time, defeating the ear. The problem is that, at normal rates, speech produces from eight to ten segments per second, and, for short stretches, at least double that number. But if each of those were a unit sound, then rates that high would strain the temporal resolving power of the ear, and, of particular importance to phonetic communication, also exceed its abililty to perceive the order in which the segments had been laid down. Indeed, the problem would be exactly the one we encountered when, in the early work on reading machines, we presented acoustic alphabets at high rates.

According to the vertical view, nature solved the rate problem by avoiding the acoustic-auditory (horizontal) strategy that would have caused it. What evolved as the phonetic constituents were the special gestures I spoke of earlier. These serve well as the elements of language, because, if properly chosen and properly controlled, they can be coarticulated, so strings of them can be produced at high rates. In any case, all speakers of all languages do, in fact, coarticulate, and it is only by this means that they are able to communicate phonologically as rapidly as they do.

Coarticulation had happy consequences for perception, too. For coarticulation folds information about several successive segments into the same stretch of sound, thereby achieving a parallel transmission of information that considerably relaxes the constraint imposed by the temporal resolving properties of the ear. But this gain came at the price of a relation between acoustic signal and phonetic message that is complex in a specifically phonetic way. One such complication is the context-conditioned variability in the acoustic signal that I identified as the primary motivation for the Early Motor Theory, presenting it then as if it were an obstacle that the processes postulated by the theory had to overcome. Now, on the Revised

Motor Theory, we can see that same variability as a blessing, a rich source of information about phonetic structure, and, especially, about order. Consider, again, the difficulty the auditory system has in perceiving accurately the ordering of discrete and brief sounds that are presented sequentially. Coarticulation effectively gets around that difficulty by permitting the listener to apprehend order in quite another way. For, given coarticulation, the production of any single segment affects the acoustic realization of neighboring segments, thereby providing, in the context-conditioned variation that results, accurate information about which gesture came first, which second, and so on. Hence, the order of the segments is conveyed largely by the shape of the acoustic signal, not by the way pieces of sound are sequenced in it. For example, in otherwise comparable consonant-vowel and vowel-consonant syllables, the listener is not likely to mistake the order, however brief the syllables, because the transitions for prevocalic and postvocalic consonants are mirror images. But these will have the proper perceptual consequence only if the phonetic system is specialized to represent the strongly contrasting acoustic signals, not as similarly contrasting auditory percepts, but as the opening and closing phases of the same phonetic gesture. Accordingly, order is given for free by processes that are specialized to deal with the acoustic consequences of coarticulated phonetic gestures. Thus, we see that a critical function of the phonetic module is not so much to take advantage of the properties of the general motor and auditory systems—a matter that was briefly examined earlier—as it is to find a way around their limitations.

Could the assignment of the stimulus information to phonetic categories plausibly be auditory?

Many of the empirically based arguments about the two theories of speech, including, especially, most of those that have been advanced against the vertical position, come from experiments, much like those described in the first section of this essay, that were designed very simply to identify the information that leads to perception of phonetic segments. The results of these experiments have proved to be reliable, so there is quite general agreement about the nature of the relevant information. Disagreement arises only, but nonetheless fundamentally, about the nature of the event that the information is informing about. Is it the sound, as a proper auditory (and horizontal) view would have it, or the articulatory gesture, which is the choice of the vertically oriented Motor Theory.

The multiplicity of acoustic-phonetic boundaries and cues. Research of the kind just referred to has succeeded in isolating many acoustic variables important to perception of the various phonetic segments, and in finding for each the location of the boundary that separates the one segment from some alternative—for example, [ba] from [da]. The horizontalists take satisfaction in further experiments on some of these boundaries in which it has been found that they are exhibited by nonhuman animals, or by human observers when presented with nonspeech analogs of speech, for these findings are, of course, consistent with the assumption that the boundaries are auditory in nature. In response, the verticalists point to experiments in which it has been found that the boundaries differ between human and nonhuman subjects, and, in humans, between speech and nonspeech analogs, arguing, in their turn, that these findings support the view that the boundaries are specifically phonetic. Indeed, for some parties to the debate it has been in the interpretation of these boundaries that the difference between the two views has come into sharpest focus. The issue therefore deserves to be further ventilated.

It is now widely accepted that the location of the acoustic-phonetic boundary on every relevant cue dimension varies greatly as a function of phonetic context, position in the syllable, and vocal-tract dimensions. It is now also known, and accepted, that some vary with differences in language type, linguistic stress, and rate of articulation. For at least one of these—rate of articulation—the variation is possibly continuous. From all this it follows that the number of acoustic-phonetic boundaries is indefinitely large, far too large, surely, to make reasonable the assumption that they are properties of the auditory system. How would these uncountably many boundaries have been selected for as that system evolved? Surely, not just against the possibility that language would come along and find them useful. Indeed, as auditory properties, they would presumably be dysfunctional, since they are perceptual discontinuities of a sort, and would, therefore, cause continuously varying acoustic events to be perceived discontinuously, thereby frustrating veridical perception.

The matter is the worse confounded for the auditory theory when proper account is taken of the fact that, for every phonetic segment, there

are multiple cues, and that phonetic perception uses all of them. For if, in accounting for the perception of certain consonantal segments, we attribute an auditory basis to all the context-variable boundaries on, say, the second-formant transitions—already a dubious assumption, as we've seen—then what do we do about the third-formant transitions and the bursts (or fricative noises)? These various information-bearing aspects of the signal are not independently controllable in speech production, so one must wonder about the probability that a gesture so managed as to have just the 'right' acoustic consequences for the second-formant transition would happen, also, to have just the right consequences in all cases for the other, acoustically very different cues. On its face, that probability would seem to be vanishingly small.

Nor does it help the horizontal position to suggest, as some have, that the acoustic-phonetic boundaries exhibited by nonhuman animals served merely as the auditory starting points—the protoboundaries, as it were—to which all the others were somehow added. For this is to suppose that, out of the many conditions known to affect the acoustic manifestation of each phonetic segment, some one is canonical. But is it plausible to suppose that there really are canonical forms for vocal-tract size, rate of articulation, condition of stress, language type, and all the other conditions that affect the speech signal? And what of the further implications? What, for example, is the status of the countless other boundaries that had then to be added in order to accommodate the noncanonical forms? Did they infect the general auditory system, or were they set apart in a distinct phonetic mode? If the former, then why does everything not sound very much like speech? If the latter, then are we to suppose that the listener shifts back and forth between auditory and phonetic modes depending on whether or not it is the canonical form that is to be perceived?

None of this is to say that natural boundaries or discontinuities do not exist in the auditory system—I believe there is evidence that they do—but rather to argue that they are irrelevant to phonetic perception.

All of the foregoing considerations are simply grist for the Motor Theory mill. For on that theory, the phonetic module uses the speech stimulus as information about the gestures, which are the true and immediate objects of phonetic perception, and so finds the acoustic-phonetic boundaries where the articulatory apparatus happened, for its own very good reasons, to put them. The auditory system is then free to respond to all other acoustic stimuli in a way that does not inappropriately conform their perception to a phonetic mold.

Integrating cues that are acoustically heterogeneous, widely distributed in time, and shared by disparate segments. Having already noted that there are typically many cues for a phonetic distinction, I take note of the well-known fact that these many cues are, more often than not, acoustically heterogeneous. Yet, in the several cases so far investigated, they can, within limits, be traded, one for the other, without any change in the immediate percept. That is, with other cues neutralized, the phonetic distinction can be produced by any one of the acoustically heterogeneous cues, with perceptual results that are not discriminably different. Since these perceptual equivalences presumably exist among all the cues for each such contrast, the number of equivalences must be very great, indeed. But how are these many equivalences to be explained? From an acoustic or auditory-processing standpoint, what do such acoustically diverse, but perceptually equivalent, cues have in common? Or, in the absence of that commonality, how plausible is it to suppose that they might nevertheless have evolved in the auditory system in connection with its nonspeech functions? In that regard, one asks the same questions I raised about the claim concerning the auditory basis of the boundary positions. What general auditory function would have selected for these equivalences? Would they not, in almost every case, be dysfunctional, since they would make very different acoustic events sound the same? And, finally, what is the probability that speakers could so govern their articulatory gestures as to produce for each particular phonetic segment exactly the right combination of perceptually equivalent cues?

The Revised Motor Theory has no difficulty dealing with the foregoing facts about stimulus equivalence. It notes simply that the acoustically heterogeneous cues have in common that they are products of the same phonetically significant gesture. Since it is the gesture that is perceived, the perceptual equivalence necessarily follows.

Also relevant to the argument is the fact that phonetic perception integrates into a coherent phonetic segment a numerous variety of cues that are, because of coarticulation, widely dispersed through the signal and used simultaneously to provide information for other segments in the string, including not only their position, as earlier noted, but also their phonetic identity. The sim-

plest examples of such dispersal were among the very earliest findings of our speech research, as described in the first half of this essay. Since then, the examples have been multiplied to include cases in which the spread of the cues for a single segment is found to be much broader than originally supposed, extending, in some utterances, from one end of a complex syllable to the other; yet, even in these cases, the phonetic system integrates the information appropriately for each of the constituent segments. I have great difficulty imagining what function such integration would serve in a system that is adapted to the perception of nonspeech events. Indeed, I should suppose that it would sometimes distort the representation of events that were discrete and closely sequenced.

Again, the Revised Motor Theory has a ready, if by now expected, explanation: the widely dispersed cues are brought together, as it were, into a single and perceptually coherent segment because they are, again, the common products of the relevant articulatory gesture.

Integrating acoustic and optical information. It is by now well known that, as Harry McGurk demonstrated some years ago, observers form phonetic percepts under conditions in which some of the information is acoustic and some optical, provided the optical information is about articulatory gestures. Thus, when observers are presented with acoustic [ba], but see a face saying [de], they will, under many conditions of intensity and clarity of the signal, perceive [da], having taken the consonant from what they saw and the vowel from what they heard. Though the perceptual effect is normally quite compelling, the result is typically experienced as slightly imperfect by comparison with the normal case in which acoustic and optical stimuli are in agreement. But the observers can't tell what the nature of the imperfection is. That is, they can't say that it is to be attributed to the fact that they heard one thing but saw another. Left standing, therefore, is the conclusion that the McGurk effect provides strong evidence for the equivalence in phonetic perception of two very different kinds of physical information, acoustic and optical.

For those who believe that speech perception is auditory, the explanation of the McGurk effect must be that the unitary percept is the result of a learned association between hearing a phonetic structure and seeing it produced. As an explanation of the phenomenon, however, such an account seems manifestly improbable, since it requires us to believe, contrary to all experience, that a convincing auditory percept can be elicited by an optical stimulus, or that an auditory percept and a visual percept become indistinguishable as a consequence of frequent association. Indeed, we are required to believe, even more implausibly, that the seemingly auditory percept elicited by the optical stimulus is so strong as to prevail over the normal (and different) auditory response to a suprathreshold *acoustic* stimulus that is presented concurrently. If there were such drastic perceptual consequences of association in the general case, then the world would sometimes be misrepresented to observers as they gained experience with percepts in different modalities that happened often to be contiguous in time. Fortunately for our relation to the world, there is no reason to suppose that such modality shifts, and the consequent distortions of reality, ever occur. As for the implications of the horizontal account for the McGurk effect specifically, we should expect that the phenomenon would be obtained between the sounds of speech and print, given the vast experience that literate adults have had in associating the one with the other. Yet the effect does not occur with print. It also weighs against the same account that prelinguistic infants have been shown to be sensitive to the correspondence between speech sounds and seen articulatory movements, which is, of course, the basis of the McGurk effect.

On the vertical view, the McGurk phenomenon is exactly what one would expect, since the acoustic and optical stimuli are providing information about the same phonetic gesture, and it is, as I have said so relentlessly, precisely the gesture that is perceived.

Just how 'special' is speech perception?

The claim that speech percep-tion is special has been criticized most broadly, perhaps, on the ground that it is manifestly unparsimonious and lacking in generality. Unparsimonious, because a "special" mechanism is necessarily an additional mechanism; and lacking in generality, because that which is special is, by definition, not general.

As for parsimony, I have already suggested that the shoe is on the other foot. For the price of denying a distinctly phonetic mode at the level of perception is having to make the still less parsimonious assumption that such a mode begins at a higher cognitive level, or wherever it is that the auditory percepts of the horizontal view are converted to the postperceptual phonetic shapes they must assume if they are to serve as the vehicles of linguistic communication.

But generality is another matter. Here, the horizontal view might appear to have the advantage,

since it sees the perception of speech as a wholly unexceptional example of the workings of an auditory modality that deals with speech just as it does with all the other sounds to which the ear is sensitive. In so doing, however, this view sacrifices what is, I think, a more important kind of generality, since it makes speech perception a mere adjunct to language, having a connection to it no less arbitrary than that which characterizes the relation of language to the visually perceived shapes of an alphabet. The vertical view, on the other hand, shows the connection to language to be truly organic, permitting us to see speech perception as special in much the same way that other components of language perception are special. I have already pointed out in this connection that the output of the specialized speech module is a representation that is, by its nature, specifically appropriate for further processing by the syntactic component. Now I would add that the processes of phonetic and syntactic perception have in common that the distinctly linguistic representations they produce are not given directly by the superficial properties of the signal. Consider, in this connection, how a perceiving system might go about deciding whether or not an acoustic signal contains phonetic information. Though there are, to be sure, certain general acoustic characteristics of natural speech, experience with synthetic speech has shown that none of them necessarily signals the presence of phonetic structure. Having already noted that this was one of the theoretically interesting conclusions of the earliest work with the the highly schematized drawings used on the Pattern Playback, I add now that more convincing evidence of the same kind has come from later research that carried the schematization of the synthetic patterns to an extreme by reducing them to three sine waves that merely follow the center frequencies of the first three formants. These bare bones have nevertheless proved sufficient to evoke phonetic percepts, even though they have no common fundamental, no common fate, nor, indeed, any other kind of acoustic commonality that might provide auditory coherence and mark the sinusoids acoustically as speech. What the sinusoids do offer the listener—indeed, all they offer—is information about the trajectories of the formants, which is to say movements of the articulators. If those movements can be seen by the phonetic module as appropriate to linguistically significant gestures, then the module, being properly engaged, integrates them into a coherent phonetic structure; otherwise, not. There are. then, no purely acoustic properties, no acoustic stigmata, on the basis of which the presence of phonetic structure can, under all circumstances, be reliably apprehended. But is it not so with syntax, too? If a perceiving system is to determine whether or not a string of words is a sentence, it surely cannot rely on some list of surface properties; rather it must determine if the string can be parsed—that is, if a grammatical derivation can be found. Thus, the specializations for phonetic and syntactic perception have in common that their products are deeply linguistic, and are arrived at by procedures that are similarly synthetic.

As for specializations that are adapted for functions other than communication, Mattingly and I have claimed for speech perception that, as I earlier hinted, it bears significant resemblances to a number of biologically coherent adaptations. Being specialized for different functions, each of these is necessarily different from every other one, but they nevertheless have certain properties in common. Thus, they all qualify as modules, in Fodor's sense of that term, and therefore share something like the properties he assigns to such devices. I choose not to review those here, but rather to identify, though only briefly, several other common properties that such modules, including the phonetic, seem to have.

To see what some of those properties might be, Mattingly and I have found it useful to distinguish broadly between two classes of modules. One comprises, in the auditory modality, the specializations for pitch, loudness, timbre, and the like. These underlie the common auditory dimensions that, in their various combinations, form the indefinitely numerous percepts by which people identify a correspondingly numerous variety of acoustic events, including, of course, many that are produced by artifacts of one sort or another, and that are, therefore, less than perfectly natural. We have thought it fitting to call this class 'open'. It is appropriate to the all-purpose character of these modules that its representations be commensurate with the relevant dimensions of the physical stimulus. Thus, pitch maps onto frequency, loudness onto amplitude, and timbre onto spectral shape; hence, we have called these representations 'homomorphic'. It is also appropriate to their all-purpose function that these homomorphic representations not be permanently changed as a result of long experience with some acoustic event. Otherwise, acquired skill in using the sound of automobile engines for diagnostic purposes, for example, would render the relevant modules maladapted for every one of the many other events for which they must be used.

Members of the other class—the one that includes speech—are more narrowly specialized for particular acoustic events or stimulus relationships that are, as particular events or relationships, of particular biological importance to the animal. We have, therefore, called this class 'closed'. It includes specializations like sound localization, stereopsis, and echolocation (in the bat) that I mentioned earlier. Unlike the representations of the open class, those produced by the closed modules are incommensurate with the dimensions of the stimulus; we have therefore called them 'heteromorphic'. Thus, the sound-localizing module represents interaural differences of time and intensity, not homomorphically as time or loudness, but heteromorphically as location; the module for stereopsis represents binocular disparities, not homomorphically as double images, but heteromorphically as depth. The echo-locating module of the bat presumably represents echo time, not homomorphically as an echoing (bat) cry, but heteromorphically as distance. In a similar way, the phonetic module respresents the continuously changing formants, not homomorphically as smoothly varying timbres, but heteromorphically as a sequence of discrete and categorical phonetic segments.

Unlike the open modules, those of the closed class depend on very particular kinds of environmental stimulation, not only for their development, but for their proper calibration. Moreover, they remain plastic—that is, open to calibration—for some considerable time. Consider, in this connection, how the sound-localizing module must be continuously recalibrated for its response to interaural differences as the distance between the ears increases with the growth of the child's head. The similarly plastic phonetic module is calibrated over a period of years by the particular phonetic environment to which it is exposed. Significantly, the calibration of these modules in no way affects any of the specializations of the open class, even though their representations figure importantly in the final percept, as in the paraphonetic aspects of speech, for example, This is to say that the closed modules must learn by experience, as the phonetic module most surely does, but the learning is of an entirely precognitive sort, requiring only neurological normality and exposure (at the right time) to the particular kinds of stimuli in which the module is exclusively interested.

As implied above, the two classes have their own characteristically different ways of representing the same dimension of the stimulus. Why, then, does the listener not get both representations—heteromorphic and homomorphic—at the same time? Given binocularly disparate stimuli, why does the viewer not see double images in addition to depth? Or, given two syllables [da] and [ga] that are distinguished only by the direction of the third-formant transition, why does the listener not hear, in addition to the discrete consonant and vowel, the continuously changing timbre that the most nearly equivalent nonspeech pattern would produce, and to which the two transitions would presumably make their distinctively different, but equally nonphonetic, contributions.

Mattingly and I have proposed that the competition between the modules is normally resolved in favor of the members of the closed class by virtue of their ability to preempt the information that is ecologically appropriate to computing the heteromophic percept, and thus, in effect, to remove that information from the flow. As for what is ecologically appropriate, the closed modules have an elasticity that permits them to take a rather broad view. Thus, the module for stereopsis represents depth for binocular disparities considerably greater than would ever be produced by even the most widely separated eyes. The phonetic module will, in its turn, tolerate rather large departures from what is ecologically plausible. Imagine, for example, two synthetic syllables, [da] and [ga], distinguished only by the direction of a third-formant transition that. as I indicated earlier, sounds in isolation like a nonspeech chirp. If the syllables are now divided into two parts—one, the critical transition cue; the other, the remainder of the pattern (the 'base') that, by itself, is ambiguously [da] or [ga]—then, the phonetic system will integrate them into a coherent [da] or [ga] syllable even when the two parts have, by various means, been made to come from perceptibly different sources. (Mattingly has a different interpretation of this particular phenomenon, so I must take full responsibility for the one I offer here,) In what is, perhaps, the most dramatic demonstration of this kind of integration, the isolated transition cue is presented at a location opposite one ear, the ambiguous base at a location opposite the other. Under these ecologically implausible circumstances, the listener nevertheless perceives a perfectly coherent [da] or [ga], and, more to the point, confidently localizes it to the side where only the ambiguous base was presented. (This happens, indeed, even when both the base and the critical transition cue are made of frequency-modulated sinusoids, which is a most severe test, since the differently located sinusoids would, as I pointed out earlier, seem to lack any kind of acoustic co-

herence that might cause them to be integrated on some auditory basis.) But there are limits to this elasticity, and seemingly similar effects occur in phonetic perception and in stereopsis when those limits are exceeded. Thus, in the speech case, as the stimulus is made to provide progressively more evidence for separate sources—for example, by increasing the relative intensity of the isolated transition on the one side—listeners begin to hear both the integrated syllable *and* the nonspeech chirp. In other words, the heteromorphic and homomorphic percepts are simultaneously represented. In the speech case, this has been called 'duplex perception'. Appropriate (and necessary) tests for the claim that the duplex percepts are representations of truly different types, and not simply a cognitive reinterpretation of a single type, are in the demonstration that listeners cannot accurately identify the chirps as [da] and [ga], and that they hear only the integrated syllable and the chirp, not those and also the ambiguous base. It is also relevant to the claim that the discrimination functions obtained for the percepts on the two sides of the percept are radically different, for that shows that the listeners cannot hear the one in the other, even when, under the conditions of the discrimination procedure, they try. (Unfortunately, these tests have not been applied to the cases of nonspeech auditory perception that some have claimed to be duplex.)

In the case of stereopsis, the elasticity of the module is strained by progressively increasing the binocular disparity. Beyond a certain point, the viewer begins to perceive, not only heteromorpic depth, but also homomorphic double images. This seems quite analogous to duplex perception, though it has not been called that.

In both speech and stereopsis, providing further evidence of ecological implausibility causes the heteromorphic percept (phonetic structure or depth) to weaken as its homomorphic counterpart (nonphonetic chirp or double images) strengthens, until, finally, the closed module fails utterly, and only the homomorphic percept of the open modules is represented. Thus, the information in the stimulus can seemingly be variously divided between the two kinds of representation, and, since either gains at the expense of the other, it is as if there were a kind of conservation of information.

Putting the matter most generally, then, I should say that speech is special, to be sure, but neither more nor less so than many other biologically coherent adaptations, including, of course, language itself.

How do speaking and listening differ from writing and reading?

Among the most obvious, and obviously consequential, facts about language is the immense difference in biological status, hence naturalness, between speech, on the one hand, and writing/reading, on the other. The phonetic units of speech are the vehicles of every language on earth, and are commanded by every neurologically normal human being. On the other hand, many, perhaps most, languages do not even have a written form, and, among those that do, some competent speakers find it all but impossible to master. Having been thus reminded once again that speech is the biologically primary form of the phonological behavior that typifies our species, we readily appreciate that alphabetic writing is not really the primary behavior itself, but only a fair description of it. Since what is being described is species typical, alphabetic writing is a piece of ethological science, in which case a writer/reader is fairly regarded as an accomplished ethologist. It weighs heavily against the horizontal view, therefore, that, as I have already said and will say again below, it cannot comprehend the difference between speaking/listening and writing/reading, for in that respect it is like a theory of bird song that does not distinguish the behavior of the bird from that of the ethologist who describes it.

To see the problem created by the horizontal view, we need first to appreciate, once again, that writing-reading did not evolve as part of the language faculty, so the relevant acts and percepts cannot be specifically linguistic. The important consequence, of course, is that they require to be made so, and, as I have said so many times, that can be done only by some kind of cognitive translation. Now I would emphasize that it is primarily in respect of this requirement that writing and reading differ biologically from speech. Indeed, it is precisely in the need to meet this requirement that writing-reading are intellectual achievements in a way that speech is not. But the horizontal view of speech does not permit us to see that essential difference. Rather, it misleads us into the belief that the primary processes of the two modes of linguistic communication are equally general, hence equally nonphonetic. That being so, we must suppose that the relevant representations are equally in need of a cognitive connection to the language, and so have the same status from a biological point of view. We are, of course, permitted to see the obvious and superficial differences, but, for each one of those, the horizontal view would

seem, paradoxically, to give the advantage to writing-reading, leading us to expect that writing-reading, not speaking-listening, would be the easier and more natural. For, surely, the printed characters offer a much better signal-to-noise ratio than the phonetically relevant parts of the speech sound; the fingers and the hand are vastly more versatile than the tongue; the eye provides far greater possibilities for the transmission of information than the ear; and, for all the vagaries of some spelling systems, the alphabetic characters bear a more nearly transparent relation to the phonological units of the language than the context-variable and elaborately overlapped cues of the acoustic signal.

The vertical view, on the other hand, is appropriately revealing. Given the phonetic module, speakers do not have to know how to spell a word in order to produce it. Indeed, they do not even have to know that it has a spelling. Speakers have only to access the word, however that is done; the module then spells it for them, automatically selecting and coordinating the appropriate gestures. Listeners are in similar case. To perceive the word, they need not puzzle out the complex and peculiarly phonetic relation between signal and the phonological message it conveys; they need only listen, again leaving all the hard work to the phonetic module. Being modular, these processes of production and perception are not available to conscious analysis, so the speakers and listeners cannot be aware of how they do what they do. Though the representations themselves *are* available to consciousness—indeed, if they were not, use of an alphabetic script would be impossible—they are already phonological in nature, hence appropriate for further linguistic processing, so the reader need not even notice them, as he would have to if, as in the case of alphabetic characters, some arbitrary connection to language had to be formed. Hence, the processes of speech, whether in production or perception, are not calculated to put the speaker's attention on the phonological units that those processes are specialized to manage.

On the basis of considerations very like those, Isabelle Liberman, Donald Shankweiler, and Ignatius Mattingly saw, more than twenty years ago, that, while awareness of phonological structure is obviously necessary for anyone who would make proper use of an alphabetic script, such awareness would not normally be a consequence of having learned to speak. Isabelle and Donald proceeded, then, to test this hypothesis with prelilerate children, finding that such children do, indeed, not know how to break a word into its constituent phonemes. That finding has now been replicated many times. Moreover, researchers at the Laboratories and elsewhere have found that the degree to which would-be readers are phonologically aware may be the best single predictor of their success in learning to read, and that training in phonological awareness has generally happy consequences for the progress in reading of those children who receive it. There is also reason to believe that, other things equal, an important cause of reading disability may be a weakness in the phonetic module. That weakness would make the phonological representations less clear than they would otherwise be, hence that much harder to bring to awareness. Indeed, there is at least a little evidence that reading-disabled childen do have, in addition to their problems with phonological awareness, some of the other characteristics that a weak phonetic module might be expected to produce. Thus, by comparison with normals, they seem to be poorer in several kinds of phonologically related performances: short-term-memory for phonological structures, but not for items of a nonlinguistic kind; perception of speech, but not of nonspeech sounds, in noise; naming of objects—that is, retrieving the appropriate phonological structures—even when they know what the objects are and what they do; and, finally, production of tongue-twisters.

The vertical view was not developed to explain the writing-reading process or the ills that so frequently attend it, but rather for all the reasons given in earlier sections of this paper. That it nevertheless offers a plausible account, while the horizontal view does not, is surely to be counted strongly in its favor.

Are there acoustic substitutes for speech?

When Frank Cooper and I set out to build a reading machine for the blind, we accepted that the answer to that question was not just 'yes', but 'yes, of course'. As I see it now, the reason for our blithe confidence was that, being unable to imagine an alternative, we could only think in what I have here described as horizontal terms. We therefore thought it obvious that speech sounds evoked normal auditory percepts, and that these were then named in honor of the various consonants and vowels so they could be used for linguistic purposes.

On the basis of our early experience with the nonspeech sounds of our reading machines, we learned the hard way that things were different from what we had thought. But it was not until

we were well into the research on speech that we began to see just how different and why. Now, drawing on all that research, we would say that the answer to the question about acoustic substitutes is, 'no', or, in the more emphatic modern form, 'no way'. The sounds produced by a reading machine for the blind will serve as well as speech only if they are of a kind that might have been made by the organs of articulation as they figure in gestures of a specifically phonetic sort. If the sounds meet that requirement, then they will engage the specialization for speech, and so qualify as proper vehicles for linguistic structures; otherwise, they will encounter all the difficulties that fatally afflicted the nonspeech sounds we worked with so many years ago.

FOOTNOTE

*This essay will appear as the introduction to a collection of papers to be published by MIT Press.

Haskins Laboratories Status Report on Speech Research
1993, SR-113, 33-40

On the Intonation of Sinusoidal Sentences: Contour and Pitch Height*

Robert E. Remez† and Philip E. Rubin

A sinusoidal replica of a sentence evokes a clear impression of intonation despite the absence of the primary acoustic correlate of intonation, the fundamental frequency. Our previous studies employed a test of differential similarity to determine that the tone analog of the first formant is a probable acoustic correlate of sinusoidal sentence intonation. Though the typical acoustic and perceptual effects of the fundamental frequency and the first formant differ greatly, our finding was anticipated by reports that harmonics of the fundamental within the *dominance region* provide the basis for impressions of pitch more generally. The frequency extent of the dominance region roughly matches the range of variability typical of the first formant. Here, we report two additional tests with sinusoidal replicas to identify the relevant physical attributes of the first formant analog that figure in the perception of intonation. These experiments determined (1) that listeners represent sinusoidal intonation as a pattern of *relative pitch changes* correlated with the frequency of the tonal replica of the first formant, and (2) that sinusoidal sentence intonation is probably a close match to the *pitch height* of the first formant tone. These findings show that some aspects of auditory pitch perception apply to the perception of intonation; and, that impressions of pitch of a multicomponent *nonharmonic* signal can be derived from the component within the dominance region.

When the pattern of formant frequency variation of a natural utterance is imparted to several time-varying sinusoids, there are two obvious perceptual consequences. First, the phonetic content of the original natural utterance is preserved, and listeners understand the sinusoidal voice to be saying the same sentence as the natural one on which the sinusoidal sentence is modeled (Remez, Rubin, Pisoni, & Carrell, 1981). Second, the quality of a sinusoidal voice is unnatural both in its timbre and its intonation (Remez et al., 1981; Remez, Rubin, Nygaard, & Howell, 1987; Remez & Rubin, in press). Our studies have attributed the intelligibility of sinusoidal sentences to the availability of time-varying phonetic information which is independent of the specific acoustic elements composing the signal (Remez & Rubin, 1983, 1990). Correspondingly, we have attributed the anomalous voice qualities to the absence of harmonic structure and broadband resonances from the short-term spectrum, and to the peculiar intonation of sinewave sentences (Remez et al., 1981; Remez & Rubin, 1984; in press). Unlike natural and synthetic speech, sinewave replicas of utterances do not present the listener with a fundamental frequency common to its tonal components. In the absence of this typical correlate of intonation, our studies revealed that the tonal analog of the first formant is a probable source of the impression of the odd sentence melody accompanying the phonetic perception of replicated utterances.

The studies of intonation that are reported here address two issues about the method of our earlier research (Remez & Rubin, 1984), producing the evidence that now permits a clear interpretation. These new tests show that impressions of sinusoidal intonation are approximate to the pattern of frequency change of the tone analog of the first

The authors gladly acknowledge the advice and editorial care of Stefanie Berns and Jennifer Pardo; and the assiduous scholarship and criticism of Eva Blumenthal. This research was supported by grants from NIDCD (00308) to R. E. Remez, and from NICHHD (01994) to Haskins Laboratories.

formant, and that the intonation of the sinusoidal sentence appears to approximate the specific pitch height of this crucial tone component.

Sinusoidal intonation

In our initial studies of the basis for the impression of intonation in sinewave sentences (Remez & Rubin, 1984), we used a matching task, on each trial of which the subject listened to a sinusoidal sentence followed by two single-tone patterns, and chose from the pair of single tones the better match to the intonation of the sinewave sentence. By examining the likeness judgments across a set of single-tone candidates, we were able to identify a best match, good matches, and poor matches to the intonation of a particular sentence. We generally found that the tone imitating the first formant was preferred to all other alternatives.

First, listeners consistently chose the tone replicating the first formant from a set of tonal candidates that also included the second and the third formant analogs of the sentence replica. The set of alternatives also contained a tone reproducing the greatest common divisor of the three concurrent tones of the sentence pattern; and a tone which presented a linguistically and acoustically plausible fundamental frequency contour for the sentence that we used. The expression of any clear preference among these alternatives suggested that subjects understood the instructions and were capable of the task that we set for them. Second, we found that the preference for the first formant analog did not depend on the fact that it had the greatest energy among the constituents of the sentence, for the first formant tone was the choice for matching the intonation regardless of the ordinal amplitude relations that we imposed among the tones reproducing the formant centers. Here, the alternatives in the matching set again were the tonal components of the sentence replica. Third, when listeners were asked to identify the intonation of a sentence pattern that included formant analogs with an additional tonal component below the frequency of Tone 1 that followed the natural fundamental frequency values of the utterance from which the tonal replica was derived, they did not select this Fø-analog as the match to intonation, but instead chose the first formant analog again. Here, the alternatives in the matching set were the tonal components of the sentence replica and the tone reproducing Fø variation. Last, listeners expressed no consistent experience of intonation—performance on the intonation matching task did not differ from chance—when requested to identify the intonation of a sinusoidal sentence replica that lacked the component analogous to the first formant.

Overall, these results suggested that the intonation of a sinusoidal replica of a sentence is correlated with attributes of the analog of the first formant. The selection of the first formant tone (Tone 1) as the match to intonation persisted even when that tone had neither the greatest power among the sinusoidal components of the sentence nor the lowest frequency. In that case, it is not surprising that the intonation of sinewave sentences seems odd. The pattern of first formant frequency variation is better correlated with the opening and closing of the jaw in cycles of syllable production than the patterns of rise and fall of intonation, which are correlated with the polysyllabic breath group. But, as evidence for a more specific claim—that sinusoidal sentence intonation is based on the frequency variation of the first-formant analog—our prior findings are equivocal in two crucial ways. First, if listeners chose Tone 1 because it reprised the pattern of sentence pitch, this choice may also have occurred if Tone 1 was simply the candidate falling closest in *pitch range* to the apparent intonation of the sentence. In fact, the report of Grunke & Pisoni (1982) suggests that this alternative hypothesis is plausible. They noted that the apparent similarity of speechlike syllable-length tone patterns was affected by average tone frequency as well as by details within the pattern; none of the tests of Remez & Rubin (1984) evaluated the possibility that similar effects occur in sentence length patterns. A second point to consider is that the single-tone alternatives used by Remez & Rubin (1984) in the matching test confounded pitch range and pitch pattern. It is necessary to observe matching preferences for Tone 1 in a study controlling pitch range differences among the test alternatives to conclude that listeners hear sinusoidal intonation as the correlate of Tone 1.

The two studies reported here also used a matching task to test the hypothesis that the acoustic correlate of intonation in sinewave sentences is the tone that follows the frequency and amplitude variation of the first formant. In both experiments we employed a set of alternative single-tone frequency patterns for subjects to use in matching their impressions of intonation of a four-tone sentence replica. In the first experiment, we composed the set of candidates to distinguish the perceptual effect of frequency contour from average frequency.

EXPERIMENT 1

While our prior studies had pointed to the tone analog of the first formant as the closest match to the intonation of a sinusoidal sentence, additional tests were needed to conclude that the pattern of this tone, and not merely its average frequency, was the basis for the likeness judgments. The test performed in Experiment 1 offers alternatives that have identical average frequency but different frequency patterns derived from tonal components of the sinusoidal sentence. If subjects fail here to exhibit a clear preference in matching intonation and single-tone pitch patterns, then we would conclude that the average frequency is well represented perceptually in the perception of intonation of a sinusoidal sentence, and the specific pattern of frequency changes less well.

Method

Subjects. Fifteen audiologically normal adults were recruited from the population of Barnard and Columbia Colleges. All were native speakers of English, and none had been tested in other experiments employing sinusoidal signals. The subjects were paid for participating.

Acoustic Test Materials. The acoustic materials used in this test consisted of four sinusoidal patterns: one four-tone sentence pattern, and three single-tone patterns, all of them generated by a sinewave synthesizer (Rubin, 1980). This synthesizer produces sinusoidal patterns defined by parameters of frequency and amplitude for each tone, updated at the rate of 10 ms per parameter frame. The initial synthesis parameters were obtained by analyzing a natural utterance, the sentence "My t.v. has a twelve inch screen," spoken by one of the authors. This utterance was recorded on audiotape in a sound-attenuating chamber and converted to digital records by a VAX-based pulse-code modulation system using a 4.5 kHz low-pass filter on input and a sampling rate of 10 kHz. At 10-ms intervals, center-frequency and amplitude values were determined for each of the three lowest oral formants and the intermittent fricative formant by the analysis technique of linear prediction (Markel & Gray, 1976). In turn, these values were used as sinewave synthesis parameters after correcting the errors typical of linear prediction estimates. A full description of sinusoidal replication of natural speech is provided by Remez et al. (1987).

The resulting sentence pattern comprised four time-varying sinusoids. Tone 1 corresponded to the first formant, Tone 2 to the second, Tone 3 to the third, and Tone 4 was used to replace the fricative formant. A spectrographic representation of this pattern is shown in Figure 1. The three single-tone patterns that were used to compose the pairs of alternatives in the matching trials were Tone 1 and frequency-transposed versions of Tone 2 and Tone 3, each a component of the sentence pattern that the subject heard at the beginning of each trial. Tone 1 was produced with the frequency values it exhibited within the sentence pattern. These three single-tone alternatives were produced with equal average power, and each had roughly the same average frequency of 320 Hz. This was accomplished by dividing the frequency parameters of Tones 2 or 3 by constant divisors throughout the pattern and then synthesizing the transposed time-varying sinusoids.

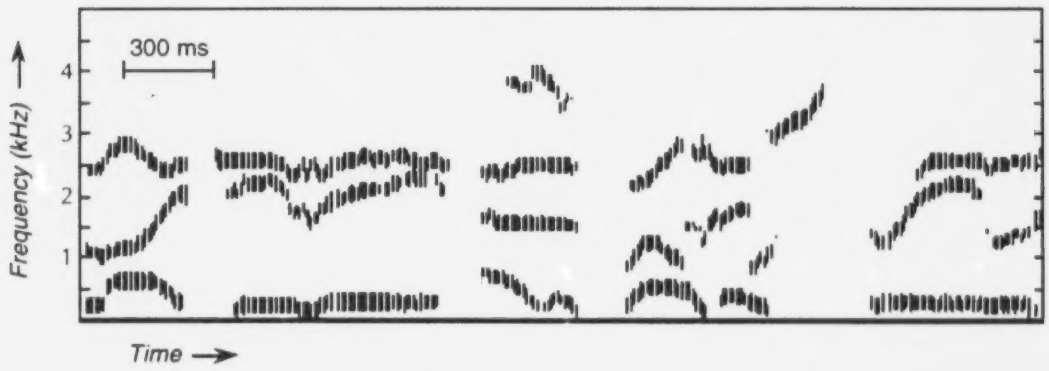


Figure 1. Spectrographic representation of the tone replica of the sentence, "My t.v. has a twelve inch screen." Amplitude variation is represented in the height of each hatch mark placed at the tone frequency.

The synthesized test materials were converted from digital records to analog signals, recorded on audiotape, and were presented to listeners by playback of the audiotape. Average listening levels were set at 72 dB SPL. Test materials were delivered binaurally in an acoustically shielded room over Telephonics TDH-39 headsets.

Procedure. During an initial instruction portion of the test session, listeners were told that the experiment was examining the identifiability of vocal pitch, the tune-like quality, of synthetic sentences. To illustrate the independence of phonetic structure and sentence melody for the test subjects, the experimenter sang the phrases "My Country 'Tis of Thee" and "I Could Have Danced All Night" with the associated melodies and with the melodies interposed. When subjects acknowledged their ability to determine the melody of a sentence regardless of its words, they were instructed to attend on each test trial to the pitch changes of the sinusoidal sentence, to identify the pattern, and then to select the alternative of the two following patterns that more closely resembled the intonation of the sentence. The subjects recorded their choices in specially prepared response booklets.

The format of each trial was identical, consisting of three sinusoidal patterns. First was the sinusoidal sentence "My t.v. has a twelve inch screen," presented once. Then, one of the three single-tone patterns was presented. Last, a second single-tone pattern was presented. There were six different comparisons among the three different single-tone alternatives. Counterbalanced for order, each subject judged each different comparison twenty times. Each sinusoidal pattern was approximately 3.1 s in duration, the interval between items within a trial was 1 s, and the interval between trials was 3 s.

Results and Discussion

The results are not difficult to interpret. Figure 2 shows the proportion of trials on which each alternative was chosen relative to the number of trials on which it was presented. Subjects preferred Tone 1 to the other two intonation candidates, and the mean differences were large. An analysis of variance was performed on the differences in preference between candidate tones in the three comparisons, and was highly significant [F (2,40)=71.62, p<.0001].

Because the single-tone alternatives differed in frequency contour but not in average frequency, the data reveal a strong preference for the first formant analog as the match to intonation, suggesting that subjects used pitch patterns rather that average frequency in this test. This finding adds credibility to our claim that the pattern of frequency variation of the first formant tone is supplying the acoustic basis for pitch impressions. Nevertheless, it remained to be shown that the precise range of frequency variation of this tone is responsible for the listener's impression of intonation in a sinusoidal sentence replica.

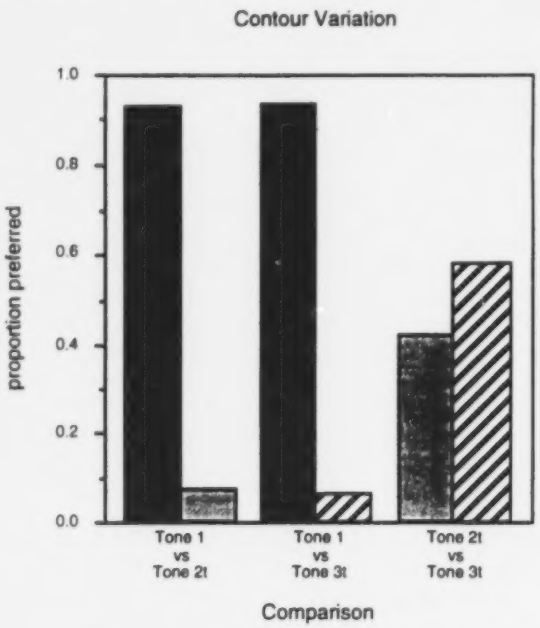


Figure 2. **Relative preference for single-tone matches to sinusoidal sentence intonation; group performance in Experiment 1. Subjects preferred Tone 1 to other tones of equal average frequency.**

EXPERIMENT 2

One way to determine the perceptual effect of the precise frequencies of the first formant analog (as opposed to the contour of its frequency variation) is to see how subjects fare in the matching task if all the alternative pitch candidates have the same contour, differing only in average frequency. If subjects persist in selecting Tone 1 as the best match when other candidates exhibit the same pattern of variation (at different pitch heights) we may conclude that Tone 1 is a fairly direct cause of pitch impressions.

In this test, we used the same sentence replica that we created for Experiment 1, but the tone alternatives for the matching task were all derived

from the pattern of the first formant analog, Tone 1. They were made by transposing the synthesis parameters for Tone 1 up or down in frequency, resulting in a set of Tone 1 variants, only one of which had the identical frequency values exhibited by Tone 1 in the sentence pattern. If subjects express precision in their preferences, we can conclude at least that the relative pitch pattern is anchored to a specific impression of pitch height, however else the impression of intonation is moderated by converging influences.

Method

Subjects. Twenty volunteer listeners with normal hearing in both ears were drawn from the student population of Barnard and Columbia Colleges. None had previously participated in studies of sinusoidal synthesis. The subjects received pay for participating.

Acoustic Test Materials. The same sinusoidal sentence pattern, "My t.v. has a twelve inch screen," from Experiment 1 was employed here. The single-tone alternatives consisted of Tone 1 and four variants: the frequency pattern of Tone 1 transposed downward by 20% and 40%, and transposed upward by 20% and 40%.

Procedure. Subjects heard the same brief instructions that were used in the first experiment to spotlight the independence of sentence melody and lexical attributes. The test itself comprised 100 trials in which each began with a single presentation of the sinusoidal sentence, "My t.v. has a twelve inch screen." Next, two of the five single tone alternatives were presented, and the subject chose the closer match to the pitch pattern of the sinusoidal sentence. There were ten different comparisons of the five alternatives, repeated five times in each order.

Results and Discussion

The outcome of this test was clear, again. An analysis of variance performed on the preference differences across the ten contests was highly significant [F (4,60)=13.79, p<.0001]. In essence, subjects consistently selected Tone 1 as most like the sentence pitch. These results are shown in Figure 3a.

The intonation matches on trials in which Tone 1 was not a candidate are shown in Figure 3b. Four of these comparisons contained a pair in which one candidate was closer to Tone 1 in frequency than the other was, and two of these comparisons had tones equally similar to Tone 1 in physical frequency. In the two equal-similarity cases, subjects chose the higher tone as the better match.

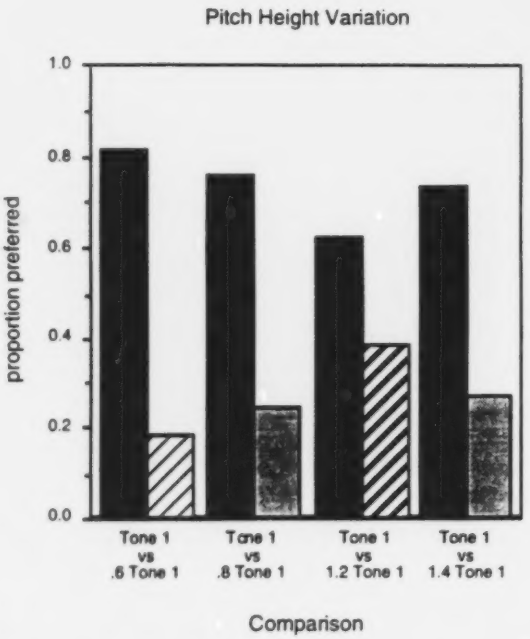


Figure 3(a). **Relative preference for single-tone matches to sinusoidal sentence intonation; group performance in Experiment 2. Conditions in which Tone 1 was compared with four transposed versions are shown. Subjects preferred Tone 1 to other tones of the same frequency contour.**

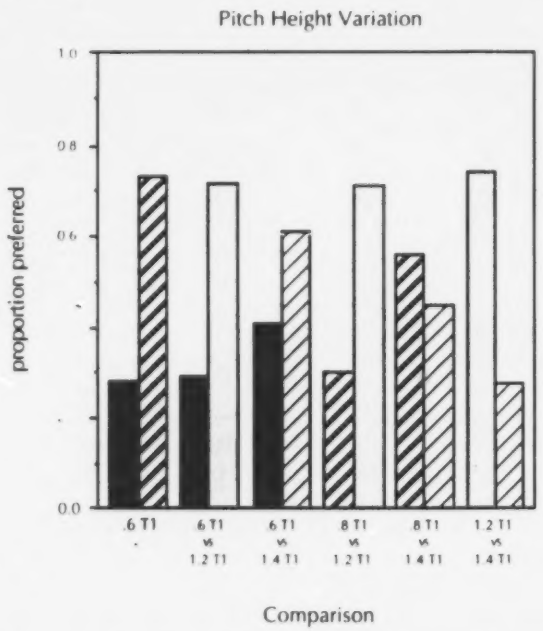


Figure 3(b). **Relative preference for single-tone matches to sinusoidal sentence intonation; group performance in Experiment 2. Conditions in which transpositions of Tone 1 were compared with each other are shown.**

This is exactly what we would expect on the precedent of classic studies of pitch scaling (cf. Ward, 1970), which roughly confirm the relationship of the 2:1 frequency ratio and the interval of the octave. In the present case, this means that the tone with its frequency increased by 40% relative to Tone 1 should be as different, subjectively, from Tone 1 as the tone with its frequency decreased 20%. This hypothesis is encouraged by the outcome of the condition comparing the case of 20% downward vs. 40% upward transpositions, in which there was no clear preference, suggesting a subjective equality of these two in degree of similarity to sentence pitch.

To summarize the ten comparisons, we derived a *likeness index* for each of the five single tone-alternatives, by summing the total number of contests over the whole test in which each was selected. Each tone occurred in forty contests, which yields a limiting score of 40 were that tone selected on every opportunity. The group averages are portrayed in Figure 4, along with the best quadratic fit to the five points. Although Tone 1 is clearly the most frequently chosen alternative across the likeness test, the effect of frequency transposition is not symmetrical.

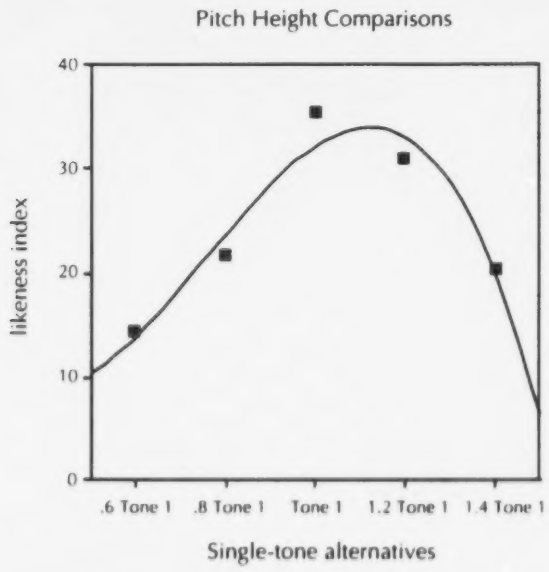


Figure 4. **Representation of the integrated performance on the ten contests of Experiment 2 is shown in this plot of overall similarity to the apparent intonation of the sinewave sentence. The likeness index for the five single-tone alternatives in Experiment 2 is graphed with a best quadratic fit to the group averages. (The curve is plotted to emphasize the asymmetry of the effect of transposition; no specific theoretical significance is given to the terms of the best-fitting function.)**

This is exactly what we would expect based on the rough match of frequency to pitch in this range, in other words, that a 40% increase in frequency is a subjectively equal change to a 20% decrease in frequency. Overall, this pattern of results suggests that the pitch height of the intonation of a sinusoidal sentence must be quite close to the pitch impression created by Tone 1 presented alone.

GENERAL DISCUSSION

The findings of Experiments 1 and 2 add precision to our account of the odd intonation that accompanies sinusoidal replicas of sentences. Listeners in our tests consistently selected the tonal analog of the first formant as the match of their impressions of intonation, as if that component of the tone complex plays a dual role in the perception of sinewave sentences: First, it acts as if it were a vocal resonance supplying segmental information about the opening and closing of the vocal tract; second, it acts as the periodic stimulation driving the perception of pitch. If the tone analog of the first formant is responsible, the intonation pattern of a sinusoidal sentence ought to sound weird, given that this tonal component (1) lies several octaves above the natural frequency of phonation, and (2) varies in a manner utterly unlike the suprasegmental pattern of Fø variation.

If the perceiver catches on to the trick of treating sinusoids as formant analogs, then it is reasonable to expect the analog of the first formant to contribute to impressions of consonants and vowels. But, why should this constituent of the signal also create an impression of intonation? These two studies of tone contour and frequency bolster the case which we have made relying on the *dominance region* hypothesis (Remez & Rubin, 1984). Research on the causes of pitch impressions with nonspeech sets of harmonic components had shown that the auditory system is roughly keyed to detect pitch preferentially from excitation in the range of 400-1000 Hz (Plomp, 1967; Ritsma, 1967). In essence, these psychoacoustic studies of fundamentals around 100-200 Hz revealed a predominant influence on pitch impressions of the common periodicity of the third through fifth harmonics, and a lesser influence of higher or lower harmonics. An important proof of the effects of the dominance region using speech sounds was reported by Greenberg (1980). This study assessed human auditory evoked potentials in response to synthetic speech spectra, and showed that the representation of the fundamental frequency in the recordings was strongest when the frequency

of the first formant fell within the dominance region. Of course, in ordinary speech, the first formant ranges widely, from roughly 270 Hz to 850 Hz in the speech of adults, to as high as 1 kHz in children. By this notion, in ordinary listening the band extending from 400-1000 Hz is analyzed to supply both periodicity information and the center-frequency of the first formant that traverses it.

The periodicity of a speech signal and the frequency of its first formant typically differ greatly in natural speech, both in pattern as well as in frequency range. In sinewave replicas, which lack harmonic structure, often the sole component falling in the dominance region is the tonal analog of the first formant. In that case, we claimed, despite the absence of harmonics, the first formant analog evidently presents an effective if inadvertent stimulus for pitch perception, and acts as well as an implicit resonant frequency. Here, the periodicity within the dominance region converges on the center frequency of the first—albeit functional—formant, and the apparent intonation is therefore unlike familiar speech in its pattern of variation and its displaced range.

The present tests furnish two missing pieces which clarify this situation. It is the frequency contour of Tone 1 which subjects use to match their impressions of sentence melody, rather than average frequency as such. Our test determined this by forcing subjects to differentiate the particular pitch contour of Tone 1 from its pitch height and pitch range. Nonetheless, subjects reported that the precise pitch height of Tone 1 matched the impression of intonation of the sentence. This was revealed in the second experiment which required subjects to differentiate frequency-transposed versions of Tone 1. Although subjects consistently chose the true Tone 1 as the best match to sentence intonation, they also confirmed, implicitly, that the familiar relation of frequency to apparent pitch obtains in this instance.

In implicating basic auditory analyzing mechanisms to account for the intonation of sinusoidal sentences, our claim suggests that definite impressions of pitch should typify tonal analogs of spoken syllables—a familiar kind of nonspeech control—whether or not phonetic segmental attributes are apparent. To take a recent instance, listeners who heard three-tone analogs of isolated steady-state vowels chose a rough approximation of the first-formant analog in matching their auditory impressions of these signals (Kuhl, Williams & Meltzoff, 1991); that result departed from the prediction that the vowels of English differed characteristically in their spectral centers of gravity, and that such integrated spectral properties ruled the perception of speech sounds. Because the listeners evidently did not perceive the tone complexes as vowels, Kuhl et al. concluded that speech and nonspeech sounds are accommodated by divergent perceptual analyses. From the perspective of our findings, though, it seems clear that pitch impressions of three-tone steady-state vowels would be derived from the frequency of the first formant analogs in any case, whether or not the spectrum of the tone complex is integrated in establishing a vowel impression. A subject who is instructed to match the vowel quality of a three-tone complex or to match its predominant pitch may attend to very different psychoacoustic attributes in each task. Although we admit that the paths to phonetic perception and auditory form perception diverge early on (Remez, Rubin, Berns, Pardo, & Lang, in press), the present studies show how intricate the interpretation of evidence can be.

While these experiments on sinusoidal sentence pitch have produced a novel instance corroborating the dominance region hypothesis, we have yet to observe a perceptual effect attributable to linguistic or paralinguistic attributes of the sentences. Some accounts of intonation report sentence-level effects based on departures from an underlying downdrift pattern, or from phrase-final fall (Vaissière, 1983). The violation of expectations based on typical properties of sentences may contribute to the apparent oddness of sinusoidal sentence melody, and to difficulties in intelligibility. But, despite the fact that tone analogs of sentences lack the familiar acoustic properties associated with a steady fundamental, we have not observed any effect of these gross departures in the registration of intonation. The pitch patterns of these sentences follow the prediction given by the dominance region hypothesis. Subsequent studies may identify precise points of departure between the impressions of intonation and the periodicity of the tone supporting the percept.

REFERENCES

Greenberg, S. (1980). Temporal neural coding of pitch and vowel quality. *Working Papers in Phonetics, 52,* 1-183. Los Angeles: U. C. L. A.

Grunke, M. E., & Pisoni, D. B. (1982). Perceptual learning of mirror image acoustic patterns. *Perception & Psychophysics, 31,* 210-218.

Kuhl, P. K., Williams, K. A., & Meltzoff, A. N. (1991). Cross-modal speech perception in adults and infants using nonspeech

auditory stimuli. *Journal of Experimental Psychology: Human Perception and Performance, 17*, 829-840.

Markel, J. D., & Gray, A. H., Jr. (1976). *Linear prediction of speech*. New York: Springer-Verlag.

Plomp, R. (1967). Pitch of complex tones. *Journal of the Acoustical Society of America, 41*, 1526-1533.

Remez, R. E., & Rubin, P. E. (1983). The stream of speech. *Scandinavian Journal of Psychology, 24*, 63-66.

Remez, R. E., & Rubin, P. E. (1984). On the perception of intonation from sinusoidal sentences. *Perception & Psychophysics, 35*, 429-440.

Remez, R. E., & Rubin, P. E. (1990). On the perception of speech from time-varying attributes: Contributions of amplitude variation. *Perception & Psychophysics, 48*, 313-325.

Remez, R. E., & Rubin, P. E. (in press). Acoustic shards, perceptual glue. In J. Charles-Luce, P. A. Luce and J. R. Sawusch (Eds.), *Theories in Spoken Language: Perception, Production, and Development*. Norwood, New Jersey: Ablex Press.

Remez, R. E., Rubin, P. E., Berns, S. M., Pardo, J. S., & Lang, J. M. (in press). On the perceptual organization of speech. *Psychological Review*.

Remez, R. E., Rubin, P. E., Nygaard, L. C., & Howell, W. A. (1987). Perceptual normalization of vowels produced by sinusoidal voices. *Journal of Experimental Psychology: Human Perception and Performance, 13*, 40-61.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell T. D. (1981). Speech perception without traditional speech cues. *Science, 212*, 947-950.

Ritsma, R. J. (1967). Frequencies dominant in the perception of the pitch of complex sounds. *Journal of the Acoustical Society of America, 42*, 191-198.

Rubin, P. E. (1980). *Sinewave synthesis*. Internal memorandum, Haskins Laboratories, New Haven, Connecticut.

Vaissière, J. (1983). Language-independent prosodic features. In A. Cutler and D. R. Ladd (Eds.), *Prosody: Models and measurements* (pp. 53-66). New York: Springer, .

Ward, W. D. (1970). Musical perception. In J. A. Tobias, (Ed.), *Foundations of modern auditory theory, Vol. 1* (pp. 407-447). New York: Academic Press.

FOOTNOTES

Journal of the Acoustical Society of America, 94, 1983-1988 (1993).

†Department of Psychology, Barnard College, New York.

Haskins Laboratories Status Report on Speech Research
1993, SR-113, 41-50

The Acquisition of Prosody: Evidence from French- and English-learning Infants*

Andrea G. Levitt†

The reduplicative babbling of five French- and five English-learning infants, recorded when the infants were between the ages of 7;3 months and 11;1 months on average, was examined for evidence of language-specific prosodic patterns. Certain fundamental frequency and syllable-timing patterns in the infants' utterances clearly reflected the influence of the ambient language. The evidence for language-specific influence on syllable amplitudes was less clear. The results are discussed in terms of a possible order of acquisition for the prosodic features of fundamental frequency, timing, and amplitude.

1. INTRODUCTION

Prosody is generally described in terms of three main suprasegmental features that vary in language-specific ways: the fundamental frequency contours, which give a language its characteristic melody; the duration or timing measures, which give a language its characteristic rhythm; and the amplitude patterns, which give a language its characteristic patterns of loud versus soft syllables. When does the prosody of infants' utterances begin to show language-specific effects?

To answer this question it is important first to understand the linguistic environment of the child, which is characterized by a special sociolinguistic register called child-directed speech (CDS). CDS has marked grammatical as well as prosodic characteristics, for which a number of possible uses have been suggested. It is also important to understand what is known about infants' sensitivity to the three prosodic features of speech. Since English and French provide very different prosodic models for young infants, they are thus excellent choices for investigating the issue of language-specific prosodic influences on infants' utterances. Analyzing the reduplicative babbling of two groups of infants, one learning French and the other English, Doug Whalen, Qi Wang and I have found evidence for the early acquisition of certain language-specific prosodic features. These results can be discussed in terms of a possible order of acquisition for language-specific prosodic features and in terms of evidence for possible regression in children's apparent sensitivity to prosodic information.

2. CHILD-DIRECTED SPEECH (CDS)

In the last twenty-five years or so, researchers have documented the existence of child-directed speech (CDS), also know as "motherese," a special style of speech or linguistic register used with young first language learners (e.g., Ferguson et al., 1986). Most researchers consider CDS to be universal (e.g., Fernald et al., 1989; but cf. Bernstein Ratner, & Pye, 1984; Heath, 1983). Compared to speech between adults, or adult-directed speech (ADS), CDS shows both special grammatical and prosodic features. From a grammatical perspective, CDS consists of shorter, simpler, and more concrete sentences, uses more repetitions, questions, and imperatives and more emphatic stress. From a prosodic perspective, CDS includes high pitch, slow rate, exaggerated pitch contours, long pauses, increased final-syllable lengthening, and whispering.

Some researchers have attributed adults' production of the higher pitch and more variable fundamental frequency of CDS to the preference of very young children for higher pitched sounds

This work was supported by NIH grant DC00403 to Catherine Best and Haskins Laboratories. We thank the families of our French and American infants for their participation in this research.

(Sachs, 1977), whereas others have focused on these prosodic characteristics as contributing to the expression of affection (Brown, 1977) or for attracting the child's attention (Garnica, 1977).

More recently, however, some investigators have argued for a more linguistically significant role for the prosodic characteristics of CDS. Thus, researchers have variously suggested that the prosodic patterns of CDS may help infants in learning how to identify their native language (Mehler et al., 1988); to identify important linguistic information, such as names for unfamiliar objects (Fernald & Mazzie, 1983); and even to parse the syntactic structures of their native language (Hirsh-Pasek et al., 1987). Some of our own current work suggests that the prosodic features of CDS may also serve to enhance speaker-specific properties of the speech signal.

As it turns out, not all of the linguistic features attributed to CDS are present at once. Indeed, certain features are quite *un*likely to co-occur. Other sociolinguistic registers remain relatively stable over time, but CDS does not. In fact, it is characterized by notable systematic changes that appear linked to the developmental stage of the child spoken to (Bernstein Ratner, 1984, 1986; Malsheen, 1980; Stern, Spieker, Barnett, & MacKain, 1983). As do the other features of CDS, the prosodic aspects also appear to change over time. For example, pitch height and the use of whispering are reduced as children grow older (Garnica, 1977). There may even be changes in the types of fundamental frequency contours that a child hears over time. A recent study (Papoušek & Hwang, 1991) has shown that Mandarin CDS prosody, as produced for presyllabic infants, may even distort the fundamental lexical tones, which are each marked by specific fundamental frequency contours in the adult language. But Chinese children do go on to learn the appropriate tones, and indeed our preliminary analyses of Mandarin CDS, produced to an infant between 9 and 11 months of age, suggest that for the older infant there is considerably less distortion. Even if very early CDS has more universal than language-specific prosodic patterns (Papoušek, Papoušek, & Symmes, 1991), the CDS addressed to older infants, as well as all other forms of speech which young infants are likely to hear, provide ample exposure to language-specific prosodic patterns as well. What is known about young infants' sensitivity to the prosodic patterns of language?

3. INFANT RESPONSE TO PROSODY

Bull and his colleagues (Bull, Eilers, & Oller, 1984, 1985; Eilers, Bull, Oller, & Lewis, 1984) have shown that infants in the second half year of life can detect changes in each of the three prosodic parameters under discussion. Researchers have found that infants' response to fundamental frequency variation or intonation is particularly strong. Indeed, infants' strong response to CDS (Fernald, 1985) can be interpreted as a preference on their part for its special fundamental frequency contours (Fernald & Kuhl, 1987). In terms of early pitch production, Kessen, Levine, and Wendrick (1979) found that infants between 3 and 6 months of age were able to match with their voices the pitches of certain notes, and there have also been reports of young infants being able to match the fundamental frequency contours of spoken utterances (Lieberman, 1986). Once children have begun to speak, they can make communicative use of pitch, e.g., contrast a request from a label, even at the one-word stage (Galligan, 1987; Marcos, 1987).

Other research has demonstrated that infants show an early perceptual sensitivity to some specific rhythmic properties of language. For example, it has been shown that very young infants can discriminate two bisyllabic utterances when they differ in syllable stress (Jusczyk & Thompson, 1978; Spring & Dale, 1977). Infants could perform this task both when the syllable stress was cued by all three typical prosodic markers as well as when the stress was cued by duration alone (Spring & Dale, 1977). Furthermore, Fowler, Smith, and Tassinary (1985) found evidence that the basis for infants' perception of speech timing is stress beats, just as it is for adults. Relatively little attention has been placed on infants' sensitivity to amplitude or loudness differences, independently of its role in stress, except for the work of Bull and his colleagues (1984), mentioned above.

4. EARLY LANGUAGE-SPECIFIC PROSODIC INFLUENCES ON PRODUCTION

Although not all attempts to find support for early language-specific effects on infant utterances have been successful (see Locke, 1983 for a review), Boysson-Bardies and her colleagues were able to find such evidence in their cross-linguistic investigations of infant utterances. For example, using acoustic analysis, they found that

the vowel formants of 10-month-old infants varied in ways consistent with the formant-frequency patterns in the adult languages (Boysson-Bardies, Halle, Sagart, & Durand, 1989). Some of our own research (Levitt & Utman, 1991), along with results from another study by Boysson-Bardies and her colleagues (Boysson-Bardies, Sagart, & Durand, 1984), suggested that young infants from different linguistic communities might also show early language-specific prosodic differences, which we decided to explore by comparing the utterances of French-learning and English-learning infants.

4.1 Prosodic differences in English and French

French and English differ in a number of ways on each of the three prosodic features. In terms of fundamental frequency contours, there are several differences, including contour shape (Delattre, 1965) and incidence of rising contours (Delattre, 1961). It is the difference in the incidence of rising contours that we investigated. Delattre (1961), who analyzed the speech of Simone de Beauvoir and Margaret Mead, found that the French speaker had many more rising continuation contours (93%) than her American counterpart (11%).

In our investigation of timing differences between French and English we focused on the syllable level, where we found at least three clear differences: the salience of final syllable lengthening, the timing of nonfinal syllables, and the interval between stressed syllables or, in other terms, the typical length of the prosodic word. In Figure 1, we can see the first two rhythmic properties illustrated for French and English. The graphs (taken from Levitt [1991]) show syllable timing measures based on reiterant productions by adult native speakers of French and English of words of two to five syllables in the two languages. The native speakers replaced the individual syllables of each word with the syllable/ma/, while preserving natural intonation and rhythm. To provide these data, ten native speakers of English and ten native speakers of French were asked to produce reiterant versions of a series of 30 words in their own language. The top graph represents timing measures for French words of two to five syllables, the middle graph represents English words of two and three syllables with all possible stress patterns, and the bottom graph represents English words of four and five syllables with a selection of stress patterns.

The first property, final-syllable lengthening, is a more salient feature of French than of English. Although both French and English exhibit final-syllable lengthening (breath-group final lengthening in French), final-syllable lengthening is more salient in French because French nonfinal syllables are not typically lengthened due to word stress, as are nonfinal syllables in English.

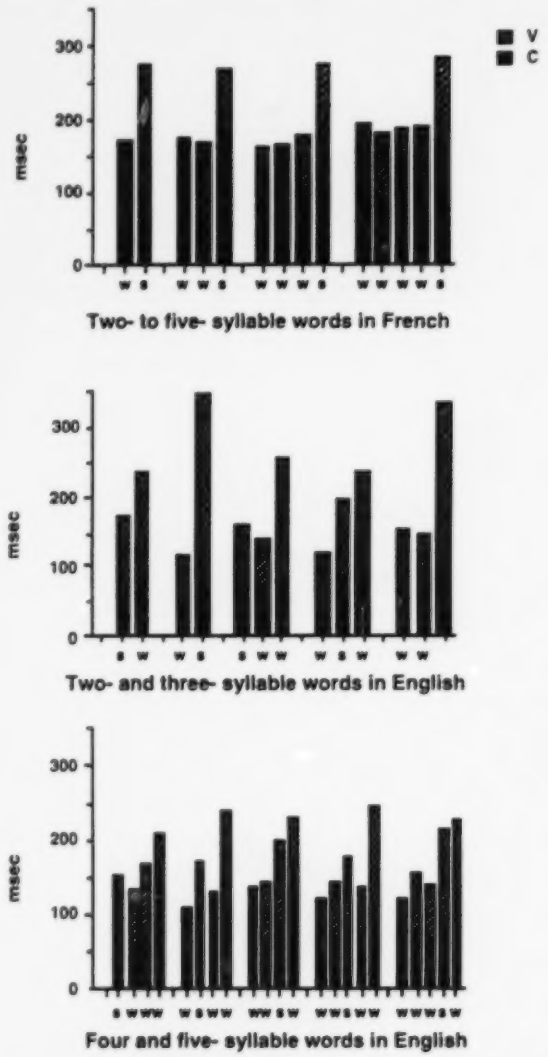


Figure 1. **Syllable timing measures for words of two- to five-syllables in French (top panel) and English (bottom two panels).**

The second properly, nonfinal syllable timing, is also clearly different for the two languages. French has been classified as syllable-timed, with a rhythmic structure known as "isosyllabicity," which is characterized by syllables generally equal in length. However, this description ignores the obvious, important final-syllable lengthening we see in French. On the other hand, aside from the effects of emphatic stress and inherent segmental

length differences, *nonfinal* syllables in French generally *are* equal in length. In Figure 1 in the top panel, the nonfinal syllables in French words of three to five syllables are quite equal in length, whereas English nonfinal syllables (in the bottom two panels) are not because of variable word stress.

Finally, the third rhythmic property that we investigated, the length of a prosodic word, here defined as the number of syllables from one stressed syllable to the next, may be expected to differ in English and French, again because of differences in the stress patterns in the two languages. Information about the typical length of the prosodic word in French and English comes from studies by Fletcher (1991) and by Crystal and House (1990). Fletcher analyzed the conversational speech of six native speakers of French. Reanalyzing a portion of her data, we found that 56% of the speakers' polysyllabic "prosodic words," which included all unaccented syllables preceding an accented final syllable, were 4 or more syllables in length, on average. On the other hand, when we examined similar data from Crystal and House, who had analyzed the read speech of six English subjects, we found that prosodic words of 4 or more syllables accounted for only six percent of the total, on average. Thus, there is some evidence that interstress intervals or prosodic words tend to be longer in French.

How do the amplitude patterns of the two languages differ? Figure 2 shows the waveforms of the French word "population" with its reiterant version, spoken by a male native speaker of French on top, and the waveforms of the English word "population" with its reiterant version, spoken by a male native speaker of English on the bottom. The patterns in Figure 2 are very representative. Basically, French words tend to start high in amplitude and generally decline, so that final syllables, which are systematically longer than nonfinal syllables, tend to be lowest in amplitude or loudness. The French reiterant version of "population," on the right, which avoids loudness variations due to inherent amplitude differences in different segments, as on the left, looks rather like a Christmas tree on its side. On the other hand, as can be seen from the waveforms for the English words, nonfinal *stressed* syllables in English tend to have greater amplitude than surrounding syllables (as well as greater duration), although there is also a tendency for the last syllable in an English word to have lower amplitude if it is not stressed.

What sorts of evidence for language-specific prosodic structure might we find in the vocal productions of young infants themselves?

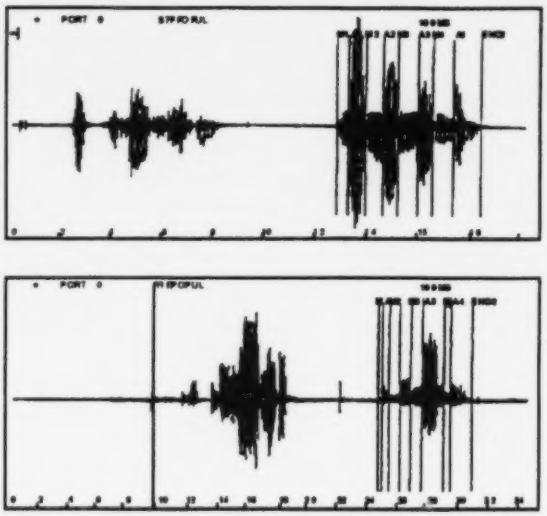


Figure 2. **Waveforms (showing characteristic amplitude patterns) of French *population* and its reiterant version (top panel) and English *population* and its reiterant version (bottom panel).**

4.2 Reduplicative babbling studies

In order to investigate whether prosodic differences in fundamental frequency contour, rhythm, or amplitude emerged in the vocal productions of French and American infants between the ages of 5 and 12 months, the babbling of five English-learning infants (three male and two female) and five French-learning infants (four male and one female) was recorded weekly by their parents at home. The French-learning infants were recorded in Paris and the English-learning infants were recorded in cities in the northeastern United States. Recordings began when the infants were between 4 and 6 months old and continued until they were between 9 and 17 months old. Each tape was phonetically transcribed, and all infant speechlike vocalizations were digitized for computer analysis. The vocalizations were divided into utterances, or breath groups, which were defined as a sequence of syllables that were separated from adjacent utterances by at least 700 ms of silence and which contained no internal silent periods longer than 450 ms in length. From the transcribed and digitized utterances, we selected all the reduplicative babbles, that is, those which contained the same consonant-like element as well

as the same vowel-like element, repeated in an utterance of two or more syllables, according to our transcriptions.

Using these criteria, we obtained 208 reduplicative utterances, approximately half (102) from the English-learning children and half (106) from the French-learning infants. (See Table 1, taken from [Levitt & Wang, 1991]).

Table 1. *Description of the source of the 208 stimuli.*

	Ages (in months) at which recordings were made	Ages (in months) at which reduplicative babbles were detected	Number of reduplicative babbles
French Infants			
MB	5-11	7-11	24
EC	6-12	6-12	42
MS	5-16	7-12	23
IZ	4-9	5- 7	9
NB	4-14	8-13	8
American Infants			
MA	5-16	8-12	24
MM	5-17	7-12	35
CR	5-17	9-10	7
AB	5-17	8-11	18
VB	4-15	7-12	18

4.2.1. Fundamental Frequency Contours. For our analysis of the fundamental frequency contours of the reduplicative babbles of the French and American infants, we decided to obtain contour judgments for the reduplicative babbles and to analyze them acoustically as well (Whalen, Levitt, & Wang, 1991). First, we asked a group of experienced listeners to judge whether each infant babble had a falling, a rising, a fall/rise, a rise/fall, or a flat contour. In order to make the perceptual judgments feasible, we limited our data set to those reduplicative babbles that were two or three syllables in length. We found both acoustic and perceptual evidence for language-specific effects in the F0 contours of the reduplicative babbles of the French- and English-learning infants.

Although about 65% of the perceptual judgments made for both the French and the American reduplicative babbles were either rise or fall, these two categories were about equally divided in the judgments of the French babbles, whereas about 75% were labelled fall for the American subjects. Thus, in agreement with the higher incidence of rising intonations in adult French speech (Delattre, 1961), the reduplicative babbles of our French infants showed a significantly higher incidence of rising F0 contours by comparison to those of our American infants.

The results of our acoustic analysis of the reduplicative babbles also supported our perceptual finding. All of the reduplicative babbles were categorized according to the contour opinion of the majority of the listeners and then acoustically analyzed. The contour patterns were then averaged for each language. The mean patterns for each of the contour types revealed an appropriate fundamental frequency curve, and statistical tests of the fundamental frequency values also support the finding that French infants produced more rising contours.

4.2.2. Timing Measures. What about timing measure differences in the infants' reduplicative babbling? We investigated this aspect of the French-learning and English-learning infants' production in another recent study (Levitt & Wang, 1991). Recall that final syllable lengthening is more salient in French, which also has more regularly timed nonfinal syllables, and longer prosodic words. Using the entire corpus of 208 utterances, we measured each syllable. A conservative criterion for measuring syllable length was adopted: the duration as measured included only the visibly voiced portion of each syllable. In order to test for final-syllable lengthening, we compared the length of the final syllables with the penultimate syllables in each utterance. To test for regularity in the timing of nonfinal syllables, we calculated the mean standard deviation of the nonfinal syllables in each utterance of three or more syllables, and to test for length of prosodic word, we looked at the number of syllables per utterance per child.

In Figure 3, the three graphs represent the results of our investigation of the timing properties. The top graph shows that the French infants had a significantly greater proportion of long final syllables (54%) than did the English-learning American infants (29%). In terms of the regularity of nonfinal syllables as shown in the middle graph, French infants produced more regular nonfinal syllables overall, although that difference was not significant.

However, when we analyzed the nonfinal syllable timing measurements in terms of an early and a late stage of reduplicative babbling production for each of the infants, we found a significant interaction, in that the nonfinal syllables of the French infants tended to become

more *regular* whereas the nonfinal syllables of the English-learning American infants tended to become more *irregular*. Finally, we also found a significant difference in the length of prosodic words, with the French infants producing considerably more longer utterances (of 4 syllables or more) than the Americans, as shown in the bottom graph of Figure 3.

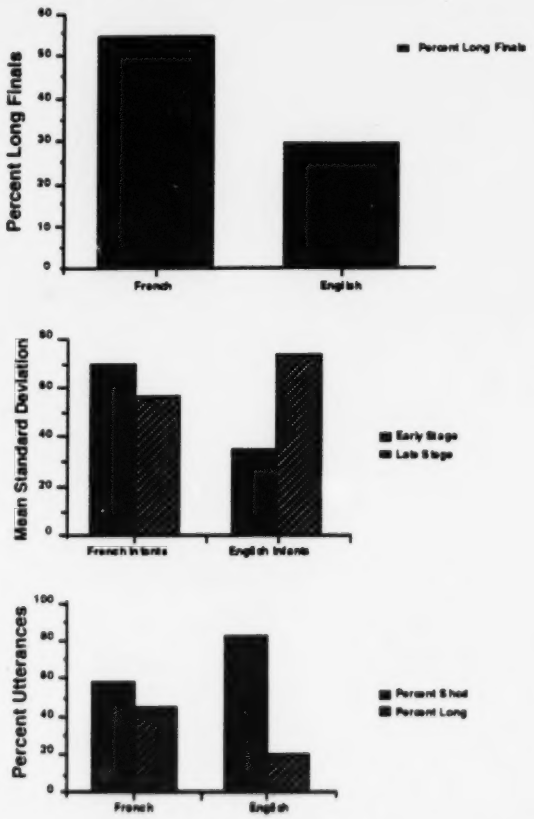


Figure 3. Comparison of French and English infants' syllable timing patterns for final-syllable lengthening (top panel) regularity of nonfinal syllables (middle panel), and number of syllables per utterance (bottom panel).

4.2.3. Amplitude Measures. What then about the last of the prosodic factors, amplitude or loudness? In order to answer this question, we first analyzed adult amplitude patterns in the two languages from the reiterant speech study mentioned earlier (Levitt, 1991). We chose five speakers of each language, 3 men and 2 women, at random. We measured the peak amplitude in each of the reiterant syllables produced by the adults and also of each of the reduplicated syllables produced by our French and American infants. Duration measures for each of the syllables had already been obtained.

Our results are pictured in the two graphs in Figure 4. As indicated in the top graph, we found that, as mentioned earlier, French adults tend to produce long final syllables with lowest amplitude (81%) significantly more often than American adults (45%) in their utterances overall [t(8)=3.2, p=.0061, one-tailed], although Americans did show a similar tendency for long finals with low amplitudes, especially for words without a final-syllable stress. The infants showed a similar pattern of results, with French infants linking long final syllables with lowest amplitude more often (33%) than English-learning American infants (21%), but this difference in the infant populations was not significant [$t(8)$=1.3, ns].

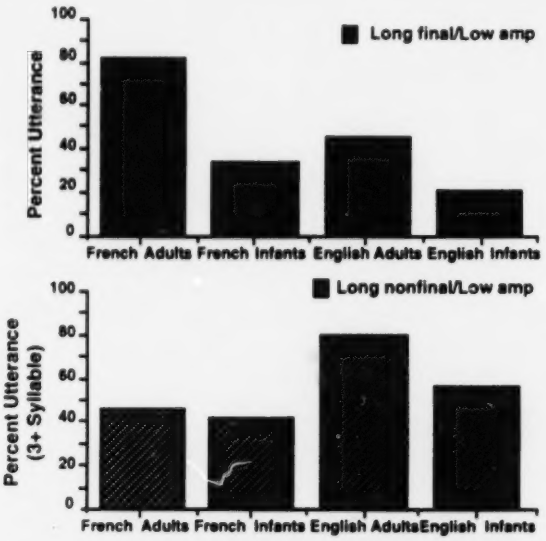


Figure 4. Comparison of duration-linked amplitude patterns for French- and English-speaking adults and French- and English-learning infants. The top panel shows the typical French pattern and the bottom panel shows the typical English pattern.

As displayed in the bottom graph of Figure 4, when we looked at nonfinal syllables in utterances of three or more syllables, we found that American adults tended to link long nonfinal syllables with *highest* amplitude or loudness (80%), significantly more often than the French adults (45%) [t(8)=4. 1, p=.0016, one-tailed]. Similarly, American infants tended to produced their highest amplitudes on the longest nonfinal syllables (57%), whereas the French infants did so less often (41%). This latter

difference between the two groups of infants approached significance [$t(8)=1.6, p=.0711$].

5. EVIDENCE FOR PROSODIC INFLUENCES IN CHILDREN'S LATER PRODUCTIONS

By the age of two, children have already largely mastered a number of the syllabic timing properties of their language. Thus, Allen (1983) has shown that French children exhibit final-syllable lengthening in polysyllabic words by two years of age. Although the patterns of final-lengthening produced by the children were more variable than those produced by French adults, the children's median ratios of final to nonfinal syllables were very comparable to those of French adults, roughly 1.6:1. Similarly, Smith (1978) has shown that English-speaking children between two and three years of age have mastered final-syllable lengthening as well, with a final to nonfinal ratio of close to 1.4:1 for both the adults and the children. Some research with two-year-old children learning tone languages (Li & Thompson, 1977) suggests that children can reproduce tonal patterns more accurately than speech segments, although certain tone contours appear easier to acquire than others.

6. POSSIBLE ORDER OF ACQUISITION OF PROSODIC FEATURES

Our results, taken together with some of the other research concerning infants' early perception (and occasional production) and young children's production of certain fundamental frequency and rhythmic properties, lend support to the notion that infants begin to imitate *some* of the prosodic properties of their native languages before they fully master its segmental properties. Specifically, it would appear that the more global properties of fundamental frequency and syllabic timing are acquired before amplitude patterns. Within each prosodic domain, there also appears to be some evidence for a learning sequence. Li and Thompson (1977), as noted above, found that children learning Mandarin acquired some tone contours, which are based on fundamental frequency, earlier than others. Similarly, our results on the acquisition of syllable timing suggest an early beginning for children's development of control of final-syllable lengthening and of utterance length, whereas acquiring the regular timing of nonfinal syllables in French appears more difficult. Children's vocal productions are notably more variable than those of adults and gradually move towards more adult-like stability as they gain increasing motor control (e.g., Kent, 1976). Producing regularly-timed syllables would thus he considerably more difficult for children than for adults.

However, before we relegate the child's control of the amplitude patterns of his/her language to the status of prosody's stepchild, we have to keep in mind that relatively little exploration has been done of the infant's sensitivity to language amplitude patterns and that the present results dealt with two languages, English and French, for which differences in the amplitude patterns of syllables may he less important in perception than are the other prosodic variables. Until more direct tests are undertaken of infants' sensitivities to all of the prosodic features and comparisons are made between languages such as English or French, on the one hand, and Ik, a language spoken in eastern Sudan, which contrasts voiceless and presumably low amplitude versus voiced, presumably high amplitude vowels, on the other hand (Maddieson, 1984), our conclusion that amplitude pattern control is acquired later than other prosodic features must be provisional.

7. EVIDENCE FOR REGRESSION IN PROSODIC LEARNING

We would also speculate, based on some of our own findings as well as on suggestions from the literature, that infants show a special sensitivity to prosody beginning perhaps at birth and lasting until about 9 or 10 months of age, when there may be some regression in the child's sensitivity to prosody. It would come, of course, at a time when Werker and her colleagues (e.g., Werker, 1989; Werker & Lalonde, 1988; Werker & Tees, 1984) and Best and her colleagues (Best, in press; Best, McRoberts, & Sithole, 1988) have shown that there is a shift in infants' phonetic perception of some nonnative segmental contrasts as their focus turns to learning the words of their native language.

Our evidence comes from both perception and production studies of prosody. Recently Catherine Best, Gerald McRoberts, and I (1991) investigated the ability of infants who were either 2-4, 6-8, or 10-12 months old to discriminate a prosodic contrast (questions versus statements) in English (their native language) and in Spanish, when there was segmental variation across the tokens representing statement and question types. The finding of interest is the comparison between the

6-8 month olds, who discriminated the prosodic contrast in both languages, and the 10-12 month olds, who failed to discriminate the questions from the statements in *both* Spanish and English. Another study, by D'Odorico and Franco (1991), which looked at infants' production of specific, prosodically-defined vocalization types in different communicative contexts, also suggested some evidence of decline toward the end of the first year. Apparently, the infants stop using these idiosyncratic, context-determined vocalizations at around 9 months of age. Finally, we also found a tendency, though not significant, for some infants in our production study to produce less consistent final syllable lengthening as they began to produce their first words (Levitt & Wang, 1991).

Although the evidence is very preliminary, the period beginning at 10 months and extending until some time before the second birthday, may be marked by some "regression" in young infants' perception and production of prosodic information.

8. CONCLUSION

It is important to remember that children learn much more from prosody than its language-specific characteristics. Prosody has a number of paralinguistic functions so that it teaches, for example, about turn taking (e.g., Schaffer, 1983) and about the expression of emotion as well (e.g., Scherer, Ladd, & Silverman, 1984). In addition to language-specific and paralinguistic function, prosody also serves some strictly grammatical linguistic functions, such as distinguishing between questions and statements or between words in a tone language. Mapping out the complete path by which children acquire the paralinguistic, language-specific, and grammatically significant prosodic patterns of their native languages, beginning from what appears to be quite an early start, has just begun.

REFERENCES

Allen, G. (1983). Some suprasegmental contours in French two-year-old children's speech. *Phonetica, 40,* 269-292.

Bernstein Ratner, N. (1984). Patterns of vowel modification in mother-child speech. *Journal of Child Language, 11,* 557-578.

Bernstein Ratner, N. (1986). Durational cues which mark clause boundaries in mother-child speech. *Journal of Phonetics, 14,* 303-309.

Bernstein Ratner, N., & Pye, C. (1984). Higher pitch in BT is not universal: Acoustic evidence from Quiche Mayan. *Journal of Child Language, 2,* 515-522.

Best, C., Levitt, A., & McRoberts, G. (1991). Examination of language-specific influences in infants' discrimination of prosodic categories. Actes du XIIème Congrès International des Sciences Phonétiques (pp. 162-165). Aix-en-Provence, France: Université de Provence Service des Publications.

Best, C., McRoberts, G., & Sithole, N. (1988). The phonological basis of perceptual loss for non-native contrasts: Maintenance of discrimination among Zulu clicks by English-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance, 14,* 345-360.

Boysson-Bardies, B. de, Sagart, L., & Durand, C. (1984). Discernible differences in the babbling of infants according to target language. *Journal of Child Language, 22,* 1-15.

Boysson-Bardies, B. de, Halle, P. Sagart, L., & Durand, C. (1989). A crosslinguistic investigation of vowel formants in babbling. *Journal of Child Language, 16,* 1-17.

Brown, R. (1977). Introduction. In C. Snow & C. Ferguson (Eds.), *Talking to children: Language input and acquisition.* Cambridge: Cambridge University Press.

Bull, D., Eilers, E., & Oller, D. (1984). Infants' discrimination of intensity variation in multisyllabic contexts. *Journal of the Acoustical Society of America, 76,* 1-13

Bull, D., Eilers, E., & Oller, D. (1985). Infants' discrimination of final syllable fundamental frequency in multisyllabic stimuli. *Journal of the Acoustical Society of America, 77,* 289-295.

Crystal, T., & House, A. (1990). Articulation rate and the duration of syllables and stress groups in connected speech. *Journal of the Acoustical Society of America, 88,* 101-112.

Delattre, P. (1961). La leçon d'intonation de Simone de Beauvoir, étude d'intonation déclarative comparée. *The French Review, 35,* 59-67.

Delattre, P. (1965). *Comparing the phonetic features of English, French, German and Spanish: An interim report.* Philadelphia: Chilton.

D'Odorico, L., & Franco, F. (1991). Selective production of vocalization types in different communicative contexts. *Journal of Child Language, 18,* 475499.

Eilers, E., Bull, D., Oller, D., & Lewis, D. (1984). The discrimination of vowel duration by infants. *Journal of the Acoustical Society of America, 75,* 213-218.

Ferguson, C. (1964). Baby talk in six languages. *American Anthropologist, 66,* 103-114.

Ferguson, C. (1978) Talking to children: A search for universals. In J. Greenberg, C. Ferguson, & E. Moravcsik (Eds.), *Universals of human language* (pp. 203-2240). Stanford: Stanford University Press.

Fernald, A. (1985). Four-month-old infants prefer to listen to motherese. *Infant behavior and development, 8,* 181 - 195.

Fernald, A., & Kuhi, P. (1987). Acoustic determinants of infant preference for motherese speech. *Infant Behavior and Development, 8,* 279-293.

Fernald, A., & Mazzie, C. (1983, April). Pitch marking of new and old information in mothers' speech. Paper presented at the meeting of the Society for Research in Child Development, Detroit.

Fernald, A., Taeschner, T., Dunn, J. Papoušek, M., Boysson-Bardies, B. de, & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language, 16,* 477-501.

Fletcher, J. (1991). Rhythm and final lengthening in French. *Journal of Phonetics, 19,* 193-212.

Fowler, C., Smith, M., & Tassinary, L. (1985). Perception of syllable timing by prebabbling infants. *Journal of the Acoustical Society of America, 79,* 814-825.

Galligan, R. (1987). Intonation with single words: Purposive and grammatical use. *Journal of Child Language, 14,* 1-21.

Garnica, O. (1977). On some prosodic and paralinguistic features of speech to young children. In C. Snow & C. Ferguson (Eds.), *Talking to children: Language input and acquisition.* Cambridge: Cambridge University Press.

Haynes, L., & Cooper, R. (1986). A note on Ferguson's proposed baby-talk universals. The Fergusonian impact: Papers in Honor

BEST COPY AVAILABLE

of Charles A. Ferguson on the Occasion of his 65th Birthday. Berlin: Mouton de Gruyter.

Heath, S. B. (1983). *Ways with words*. Cambridge: Cambridge University Press.

Hirsh-Pasek, K. Kemler-Nelson, D. G ., Jusczyk, P. W., Wright, K., Druss, B., & Kennedy, L. (1987). Clauses are perceptual units for young infants. *Cognition, 26*, 269-286.

Jusczyk, P., & Thompson, E. (1978). Perception of a phonetic contrast in multisyllabic utterances by 2-month-old infants. *Perception and Psychophysics, 23*, 105-109.

Kent, R. (1976). Anatomical and neuromuscular maturation of the speech mechanism: Evidence from acoustic studies. *Journal of Speech and Hearing Research, 19*, 421-445.

Kessen, W., Levine, J., & Wendrick, K. (1979). The imitation of pitch in infants. *Infant Behavior and Development, 2*, 93-100.

Levitt, A. (1991). Reiterant speech as a test of nonnative speakers' mastery of the timing of French. *Journal of the Acoustical Society of America, 90*, 3008-3018.

Levitt, A., & Utman, J. (1991). From babbling towards the sound systems of English and French: A longitudinal two-case study. *Journal of Child Language, 19*, 19-49.

Levitt, A., & Wang, Q. (1991). Evidence for language-specific rhythmic influences in the reduplicative babbling of French- and English-learning infants. *Language and Speech, 34*, 235-249.

Li, C., & Thompson, S. (1977). The acquisition of tone in Mandarin-speaking children. *Journal of Child Language, 4*, 185-199.

Locke, J. L. (1983). *Phonological acquisition and change*. New York: Academic Press.

Maddieson, I. (1984). *Patterns of sounds*. New York: Cambridge University Press.

Malsheen, B. (1980). Two hypotheses for phonetic clarification in the speech of mothers to children. In G. Yeni-Komshian, J. F. Kavanagh, & C.A. Ferguson (Eds.), *Child phonology: Vol. 1. Perception*. New York: Academic Press.

Marcos, H. (1987). Communicative functions of pitch range and pitch direction in infants. *Journal of Child Language, 14*, 255-268.

Mehler, J., Jusczyk. P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition, 29*, 143-178.

Papoušek, M., & Hwang, S.-F. (1991). Tone and intonation in Mandarin babytalk to presyllabic infants: Comparison with registers of adult conversation and foreign language instruction. *Applied Psycholinguistics, 12*, 481-504.

Papoušek, M., Papoušek, H., & Symmes, D., (1991). The meaning of melodies in motherese in tone and stress languages. *Infant Behavior and Development, 14*, 415.1440.

Sachs, J. (1977). The adaptive significance of linguistic input to prelinguistic infants. In C. Snow & C. Ferguson, (Eds.), *Talking to children: Language input and acquisition*. Cambridge: Cambridge University Press.

Schaffer, D. (1983). The role of intonation as a cue to turn taking in conversation. *Journal of Phonetics, 11*, 243-257.

Scherer, K., Ladd, D., & Silverman, K. (1984). Vocal cues to speaker affect: Testing two models. *Journal of the Acoustical Society of America, 76*, 1346-1356.

Smith, B. (1978). Temporal aspects of English speech production: A developmental perspective. *Journal of Phonetics, 6*, 37-67.

Spring, D , & Dale, P. (1977). The discrimination of linguistic stress in early infancy. *Journal of Speech and Hearing Research, 20*, 224-231.

Stern, D. N., Spieker, S., Barnett, R. K., & MacKain, K. (1983). The prosody of maternal speech: Infant age and context-related changes. *Journal of Child Language, 10*, 1-15.

Werker, J. F. (1989). Becoming a native listener. *American Scientist, 77*, 54-59.

Werker, J. F., & Lalonde, C. E. (1988). Cross-language speech perception: Initial capabilities and developmental change. *Developmental Psychology, 24*, 672-683.

Werker, J., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development, 7*, 49-63.

Whalen, D. H., Levitt, A., & Wang, Q. (1991). Intonational differences between the reduplicative babbling of French- and English-learning infants. *Journal of Child Language, 18*, 501-516.

FOOTNOTES

*Appears in Proceedings of the NATO Advanced Research Workshop on Changes in Speech and Face Processing in Infancy: A Glimpse at Developmental Mechanisms of Cognition. Carry-le-Rouet, France, June 29-July 3, 1992. Dordrecht, The Netherlands: Kluwer Academic Publishers.

†Also Wellesley College, Wellesley.

Haskins Laboratories Status Report on Speech Research
1993, SR-113, 51-62

Dynamics and Articulatory Phonology*

Catherine P. Browman and Louis Goldstein†

1. INTRODUCTION

Traditionally, the study of human speech and its patterning has been approached in two different ways. One way has been to consider it as mechanical or bio-mechanical activity (e.g., of articulators or air molecules or cochlear hair cells) that changes continuously in time. The other way has been to consider it as a linguistic (or cognitive) structure consisting of a sequence of elements chosen from a closed inventory. Development of the tools required to describe speech in one or the other of these approaches has proceeded largely in parallel, with one hardly informing the other at all (some notable exceptions are discussed below). As a result, speech has been seen as having two structures, one considered physical, and the other cognitive, where the relation between the two structures is generally not an intrinsic part of either description. From this perspective, a complete picture requires 'translating' between the intrinsically incommensurate domains (as argued by Fowler, Rubin, Remez, & Turvey, 1980).

The research we have been pursuing (Browman & Goldstein, 1986; 1989; 1990a,b; 1992) ('articulatory phonology') begins with the very different assumption that these apparently different domains are, in fact, the low and high dimensional descriptions of a single (complex) system. Crucial to this approach is identification of phonological units with dynamically specified units of articulatory action, called *gestures*. Thus, an utterance is described as an act that can be decomposed into a small number of primitive units (a low dimensional description), in a particular spatio-temporal configuration. The same description also provides an intrinsic specification of the high dimensional properties of the act (its various mechanical and bio-mechanical consequences).

In this chapter, we will briefly examine the nature of the low and high dimensional descriptions of speech, and contrast the dynamical perspective that unifies these to other approaches in which they are separated as properties of mind and body. We will then review some of the basic assumptions and results of developing a specific model incorporating dynamical units, and illustrate how it provides both low and high dimensional descriptions.

2. DIMENSIONALITY OF DESCRIPTION

Human speech events can be seen as quite complex, in the sense that an individual utterance follows a continuous trajectory through a space defined by a large number of potential degrees of freedom, or dimensions. This is true whether the dimensions are neural, articulatory, acoustic, aerodynamic, auditory, or other ways of describing the event. The fundamental insight of phonology, however, is that the pronunciation of the words in a given language may differ from (that is, contrast with) each other in only a restricted number of ways: the number of degrees of freedom actually employed in this contrastive behavior is far fewer than the number that is mechanically available. This insight has taken the form of the hypothesis that words can be decomposed into a small number of primitive units (usually far fewer than one hundred in a given language) which can combine in different ways to form the large number of words required in human lexicons. Thus, as argued by Kelso, Saltzman, and Tuller (1986), human speech is characterized not only by a high number of potential (microscopic) degrees of freedom, but also by a low dimensional (macroscopic) form. This macroscopic form is

This work was supported by NSF grant DBS-9112198 and NIH grants HD-01994 and DC-00121 to Haskins Laboratories. Thanks to Alice Faber and Jeff Shaw for comments on an earlier version.

usually called the 'phonological' form. As will be suggested below, this collapse of degrees of freedom can possibly be understood as an instance of the kind of self-organization found in other complex systems in nature (Haken, 1977; Kauffmann, 1991; Kugler & Turvey, 1987; Madore & Freedman, 1987; Schoner & Kelso, 1988).

Historically, however, the gross differences between the macroscopic and microscopic scales of description have led researchers to ignore one or the other description, or to assert its irrelevance, and hence to generally separate the cognitive and the physical. Anderson (1974) describes how the development of tools in the nineteenth and early twentieth centuries led to the quantification of more and more details of the speech signal, but "with such increasingly precise description, however, came the realization that much of it was irrelevant to the central tasks of linguistic science" (p.4). Indeed, the development of many early phonological theories (e.g., those of Saussure, Trubetzkoy, Sapir, Bloomfield) proceeded largely without any substantive investigation of the measurable properties of the speech event at all (although Anderson notes Bloomfield's insistence that the smallest phonological units must ultimately be defined in terms of some measurable properties of the speech signal). In general, what was seen as important about phonological units was their *function*, their ability to distinguish utterances.

A particularly telling insight into this view of the lack of relation between the phonological and physical descriptions can be seen in Hockett's (1955) familiar Easter egg analogy. The structure serving to distinguish utterances (for Hockett, a sequence of letter-sized phonological units called phonemes) was viewed as a row of colored, but unboiled, easter eggs on a moving belt. The physical structure (for Hockett, the acoustic signal) was imagined to be the result of running the belt through a wringer, effectively smashing the eggs and intermixing them. It is quite striking that, in this analogy, the cognitive structure of the speech event cannot be seen in the gooey mess itself. For Hockett, the only way the hearer can respond to the event is to infer (on the basis of obscured evidence, and knowledge of possible egg sequences) what sequence of eggs might have been responsible for the mess. It is clear that in this view, the relation between cognitive and physical descriptions is neither systematic nor particularly interesting. The descriptions share color as an important attribute, but beyond that there is little relation.

A major approach that did take seriously the goal of unifying the cognitive and physical aspects of speech description was that in the *Sound Pattern of English* (Chomsky & Halle, 1968), including the associated work on the development of the theory of distinctive features (Jakobson, Fant, & Halle, 1951) and the quantal relations that underlie them (Stevens, 1972, 1989). In this approach, an utterance is assigned two representations: a 'phonological' one, whose goal is to describe how the utterance functions with respect to contrast and patterns of alternation, and a 'phonetic' one, whose goal is to account for the grammatically determined physical properties of the utterance. Crucially, however, the relation between the representations is quite constrained: both descriptions employ exactly the same set of dimensions (the features). The phonological representation is coarser in that features may take on only binary values, while the phonetic representation is more fine-grained, with the features having scalar values. However, a principled relation between the binary values and the scales is also provided: Stevens' quantal theory attempts to show how the potential continuum of scalar feature values can be intrinsically partitioned into categorical regions, when the mapping from articulatory dimensions to auditory properties is considered. Further, the existence of such quantal relations is used to explain why languages employ these particular features in the first place.

Problems raised with this approach to speech description soon led to its abandonment, however. One problem is that its phonetic representations were shown to be inadequate to capture certain systematic physical differences between utterances in different languages (Keating, 1985; Ladefoged, 1980; Port, 1981). The scales used in the phonetic representations are themselves of reduced dimensionality, when compared to a complete physical description of utterances. Chomsky and Halle hypothesized that such further details could be supplied by universal rules. However, the above authors (also Browman & Goldstein, 1986) argued that this would not work—the same phonetic representation (in the Chomsky and Halle sense) can have different physical properties in different languages. Thus, more of the physical detail (and particularly details having to do with timing) would have to be specified as part of the description of a particular language. Ladefoged's (1980) argument cut even deeper. He argued that there is a system of scales that is useful for characterizing the measurable articulatory and acoustic properties of utterances,

but that these scales are very different from the features proposed by Chomsky and Halle.

One response to these failings has been to hypothesize that descriptions of speech should include, in addition to phonological rules of the usual sort, rules that take (cognitive) phonological representations as input and convert them to physical parameterizations of various sorts. These rules have been described as rules of 'phonetic implementation' (e.g., Keating, 1985; Keating, 1990; Klatt, 1976; Liberman & Pierrehumbert, 1984; Pierrehumbert, 1990; Port, 1981). Note that in this view, the description of speech is divided into two separate domains, involving distinct types of representations: the phonological or cognitive structure and the phonetic or physical structure. This explicit partitioning of the speech side of linguistic structure into separate phonetic and phonological components which employ distinct data types that are related to one another only through rules of phonetic implementation (or 'interpretation') has stimulated a good deal of research (e.g., Cohn, 1990; Coleman, 1992; Fourakis & Port, 1986; Keating, 1988; Liberman & Pierrehumbert, 1984). However, there is a major price to be paid for drawing such a strict separation: it becomes very easy to view phonetic and phonological (physical and cognitive) structures as essentially independent of one another, with no interaction or mutual constraint. As Clements (1992) describes the problem: "The result is that the relation between the phonological and phonetic components is quite unconstrained. Since there is little resemblance between them, it does not matter very much for the purposes of phonetic interpretation what the form of the phonological input is; virtually any phonological description can serve its purposes equally well. (p. 192)" Yet, there is a constrained relation between the cognitive and physical structures of speech, which is what drove the development of feature theory in the first place.

In our view, the relation between the physical and cognitive, i.e. phonetic and phonological, aspects of speech is inherently constrained by their being simply two levels of description—the microscopic and macroscopic—of the same system. Moreover, we have argued that the relation between microscopic and macroscopic properties of speech is one of *mutual* or *reciprocal* constraint (Browman & Goldstein, 1990b). As we elaborated there, the existence of such reciprocity is supported by two different lines of research. One line has attempted to show how the macroscopic properties of contrast and combination of phonological units arise from, or are constrained by, the microscopic, i.e., the detailed properties of speech articulation and the relations among speech articulation, aerodynamics, acoustics, and audition (e.g., Lindblom, MacNeillage, & Studdert-Kennedy, 1983; Ohala, 1983; Stevens, 1972; 1989). A second line has shown that there are constraints running in the opposite direction, such that the (microscopic) detailed articulatory or acoustic properties of particular phonological units are determined, in part, by the macroscopic system of contrast and combination found in a particular language (e.g., Keating, 1990; Ladefoged, 1982; Manuel & Krakow, 1984; Wood, 1982). The apparent existence of this bi-directionality is of considerable interest, because recent studies of the generic properties of complex physical systems have demonstrated that reciprocal constraint between macroscopic and microscopic scales is a hallmark of systems displaying 'self-organization' (Kugler & Turvey, 1987; see also discussions by Langton in Lewin, 1992 [pp. 12-14; 188-191], and work on the emergent properties of "co-evolving" complex systems: Hogeweg, 1989; Kauffman, 1989; Kauffman & Johnsen, 1991; Packard, 1989).

Such self-organizing systems (hypothesized as underlying such diverse phenomena as the construction of insect nests and evolutionary and ecological dynamics) display the property that the 'local' interactions among a large number of microscopic system components can lead to emergent patterns of 'global' organization and order. The emergent global organization also places constraints on the components and their local interactions. Thus, self-organization provides a principled linkage between descriptions of different dimensionality of the same system: the high-dimensional description (with many degrees of freedom) of the local interactions and the low-dimensional description (with few degrees of freedom) of the emergent global patterns. From this point of view, then, speech can be viewed as a single complex system (with low-dimensional macroscopic and high-dimensional microscopic properties) rather than as two distinct components.

A different recent attempt to articulate the nature of the constraints holding between the cognitive and physical structures can be found in Pierrehumbert (1990), in which the relation between the structures is argued to be a 'semantic' one, parallel to the relation that obtains between concepts and their real world denotations. In this view, macroscopic structure is constrained by the microscopic properties of speech and by the prin-

ciples guiding human cognitive category-formation. However, the view fails to account for the apparent bi-directionality of the constraints. That is, there is no possibility of constraining the microscopic properties of speech by its macroscopic properties in this view. (For a discussion of possible limitations to a dynamic approach to phonology, see Pierrehumbert & Pierrehumbert, 1990).

The 'articulatory phonology' that we have been developing (e.g., Browman & Goldstein, 1986, 1989, 1992) attempts to understand phonology (the cognitive) as the low-dimensional macroscopic description of a physical system. In this work, rather than rejecting Chomsky and Halle's constrained relation between the physical and cognitive, as the phonetic implementation approaches have done, we have, if anything, increased the hypothesized tightness of that relation by using the concept of different dimensionality. We have surmised that the problem with the program proposed by Chomsky and Halle was instead in their choice of the elementary units of the system. In particular, we have argued that it is wrong to assume that the elementary units are (1) static, (2) neutral between articulation and acoustics, and (3) arranged in non-overlapping chunks. Assumptions (1) and (3) have been argued against by Fowler et al. (1980), and (3) has also been rejected by most of the work in 'nonlinear' phonology over the past 15 years. Assumption (2) has been, at least partially, rejected in the 'active articulator' version of 'feature geometry'—Halle (1982), Sagey (1986), McCarthy (1988).

3. GESTURES

Articulatory phonology takes seriously the view that the units of speech production are actions, and therefore that (1) they are dynamic, not static. Further, since articulatory phonology considers phonological functions such as contrast to be low-dimensional, macroscopic descriptions of such actions, the basic units are (2) not neutral between articulation and acoustics, but rather are articulatory in nature. Thus, in articulatory phonology, the basic phonological unit is the articulatory gesture, which is defined as a dynamical system specified with a characteristic set of parameter values (see Saltzman, in press). Finally, because the tasks are distributed across the various articulator sets of the vocal tract (the lips, tongue, glottis, velum, etc.), an utterance is modeled as an ensemble, or constellation, of a small number of (3) potentially overlapping gestural units.

As will be elaborated below, contrast among utterances can be defined in terms of these gestural constellations. Thus, these structures can capture the low-dimensional properties of utterances. In addition, because each gesture is defined as a dynamical system, no rules of implementation are required to characterize the high-dimensional properties of the utterance. A time-varying pattern of articulator motion (and its resulting acoustic consequences) is lawfully entailed by the dynamical systems themselves—they are self-implementing. Moreover, these time-varying patterns automatically display the property of context dependence (which is ubiquitous in the high dimensional description of speech) even though the gestures are defined in a context-independent fashion. The nature of the articulatory dimensions along which the individual dynamical units are defined allows this context dependence to emerge lawfully.

The articulatory phonology approach has been incorporated into a computational system being developed at Haskins Laboratories (Browman & Goldstein, 1990a,c; Browman, Goldstein, Kelso, Rubin, & Saltzman, 1984; Saltzman, 1986; Saltzman, & Munhall, 1989). In this system, illustrated in Figure 1, utterances are organized ensembles (or *constellations*) of units of articulatory action called *gestures*. Each gesture is modeled as a dynamical system that characterizes the formation (and release) of a local constriction within the vocal tract (the gesture's functional goal or 'task'). For example, the word "ban" begins with a gesture whose task is lip closure.

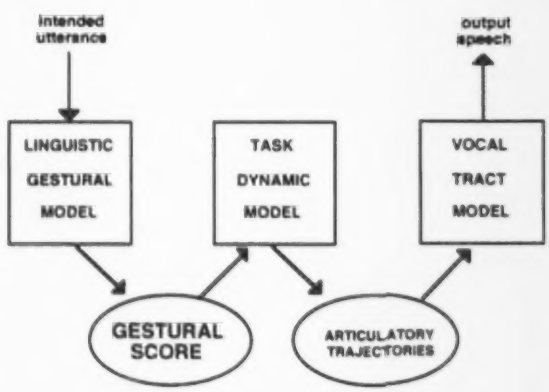


Figure 1. Computational system for generating speech using dynamically-defined articulatory gestures.

The formation of this constriction entails a change in the distance between upper and lower lips (or

Lip Aperture) over time. This change is modeled using a second order system (a 'point attractor,' Abraham & Shaw, 1982), specified with particular values for the equilibrium position and stiffness parameters. (Damping is, for the most part, assumed to be critical, so that the system approaches its equilibrium position and doesn't overshoot it). During the activation interval for this gesture, the equilibrium position for Lip Aperture is set to the goal value for lip closure; the stiffness setting, combined with the damping, determines the amount of time it will take for the system to get close to the goal of lip closure.

The set of task or *tract* variables currently implemented in the computational model are listed at the top left of Figure 2, and the sagittal vocal tract shape below illustrates their geometric definitions. This set of tract variables is hypothesized to be sufficient for characterizing most of the gestures of English (exceptions involve the details of characteristic shaping of constrictions, see Browman & Goldstein, 1989). For oral gestures, two paired tract variable regimes are specified, one controlling the constriction degree of a particular structure, the other its constriction location (a tract variable regime consists of a set of values for the dynamic parameters of stiffness, equilibrium position, and damping ratio). Thus, the specification for an oral gesture includes an equilibrium position, or goal, for each of two tract variables, as well as a stiffness (which is currently yoked across the two tract variables). Each functional goal for a gesture is achieved by the coordinated action of a set of articulators, that is, a coordinative structure (Fowler et al., 1980; Kelso, Saltzman, & Tuller, 1986; Saltzman, 1986; Turvey, 1977); the sets of articulators used for each of the tract variables are shown on the top right of Figure 2, with the articulators indicated on the outline of the vocal tract model below. Note that the same articulators are shared by both of the paired oral tract variables, so that altogether there are five distinct articulator sets, or coordinative structure types, in the system.

	tract variable	articulators involved
LP	lip protrusion	upper & lower lips, jaw
LA	lip aperture	upper & lower lips, jaw
TTCL	tongue tip constrict location	tongue tip, tongue body, jaw
TTCD	tongue tip constrict degree	tongue tip, tongue body, jaw
TBCL	tongue body constrict location	tongue body, jaw
TBCD	tongue body constrict degree	tongue body, jaw
VEL	velic aperture	velum
GLO	glottal aperture	glottis

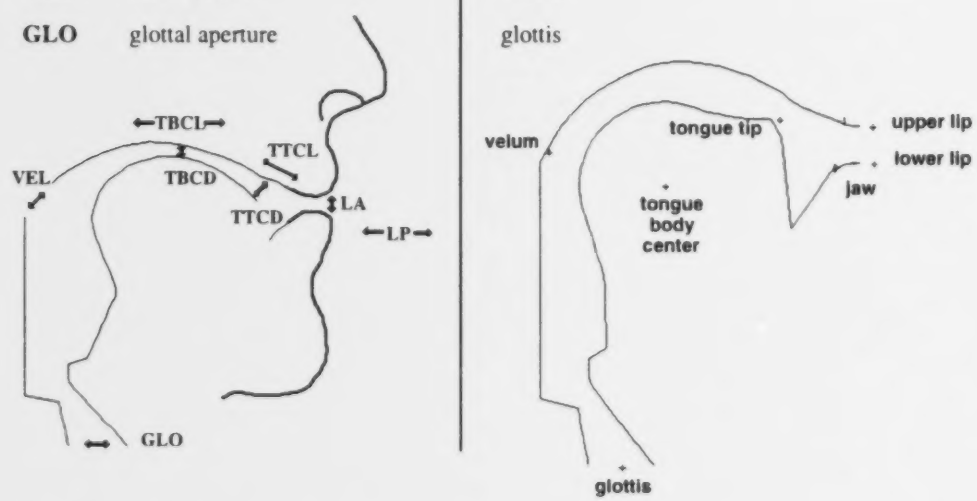


Figure 2. Tract variables and their associated articulators.

In the computational system the articulators are those of a vocal tract model (Rubin, Baer, & Mermelstein, 1981) that can generate speech waveforms from a specification of the positions of individual articulators. When a dynamical system (or pair of them) corresponding to a particular gesture is imposed on the vocal tract, the task-dynamic model (Saltzman, in press; Saltzman, 1986; Saltzman & Kelso, 1987; Saltzman & Munhall, 1989) calculates the time-varying trajectories of the individual articulators comprising that coordinative structure, based on the information about values of the dynamic parameters, etc, contained in its input. These articulator trajectories are input to the vocal tract model which then calculates the resulting global vocal tract shape, area function, transfer function, and speech waveform (see Figure 1).

Defining gestures dynamically can provide a principled link between macroscopic and microscopic properties of speech. To illustrate some of the ways in which this is true, consider the example of lip closure. The values of the dynamic parameters associated with a lip closure gesture are macroscopic properties that define it as a phonological unit and allow it to contrast with other gestures such as the narrowing gesture for [w]. These values are definitional, and remain invariant as long as the gesture is active. At the same time, however, the gesture intrinsically specifies the (microscopic) patterns of continuous change that the lips can exhibit over time. These changes emerge as the lawful consequences of the dynamical system, its parameters, and the initial conditions. Thus, dynamically defined gestures provide a lawful link between macroscopic and microscopic properties.

While tract variable goals are specified numerically, and in principle could take on any real value, the actual values used to specify the gestures of English in the model cluster in narrow ranges that correspond to contrastive categories: for example, in the case of constriction degree, different ranges are found for gestures that correspond to what are usually referred to as stops, fricatives and approximants. Thus, paradigmatic comparison (or a density distribution) of the numerical specifications of all English gestures would reveal a macroscopic structure of contrastive categories. The existence of such narrow ranges is predicted by approaches such as the quantal theory (e.g., Stevens, 1989) and the theory of adaptive dispersion (e.g., Lindblom, MacNeilage, & Studdert-Kennedy, 1983), although the dimensions investigated in those approaches are not identical to the tract variable dimensions. These approaches can be seen as accounting for how microscopic continua are partitioned into a small number of macroscopic categories.

The physical properties of a given phonological unit vary considerably depending on its context (e.g., Kent & Minifie, 1977; Öhman, 1966; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). Much of this context dependence emerges lawfully from the use of task dynamics. An example of this kind of context dependence in lip closure gestures can be seen in the fact that the three independent articulators that can contribute to closing the lips (upper lip, lower lip, and jaw) do so to different extents as a function of the vowel environment in which the lip closure is produced (Macchi, 1988; Sussman, MacNeilage, & Hanson, 1973). The value of lip aperture achieved, however, remains relatively invariant no matter what the vowel context. In the task-dynamic model, the articulator variation results automatically from the fact that the lip closure gesture is modeled as a coordinative structure that links the movements of the three articulators in achieving the lip closure task. The gesture is specified invariantly in terms of the tract variable of lip aperture, but the closing action is distributed across component articulators in a context-dependent way. For example, in an utterance like [ibi], the lip closure is produced concurrently with the tongue gesture for a high front vowel. This vowel gesture will tend to raise the jaw, and thus, less activity of the upper and lower lips will be required to effect the lip closure goal than in an utterance like [aba]. These microscopic variations emerge lawfully from the task dynamic specification of the gestures, combined with the fact of overlap (Kelso, Saltzman, & Tuller, 1986; Saltzman & Munhall, 1989).

4. GESTURAL STRUCTURES

During the act of talking, more than one gesture is activated, sometimes sequentially and sometimes in an overlapping fashion. Recurrent patterns of gestures are considered to be organized into gestural constellations. In the computational model (see Figure 1), the linguistic gestural model determines the relevant constellations for any arbitrary input utterance, including the *phasing* of the gestures. That is, a constellation of gestures is a set of gestures that are coordinated with one another by means of phasing, where for this purpose (and this purpose only), the dynamical regime for each gesture is treated as if it were a cycle of an

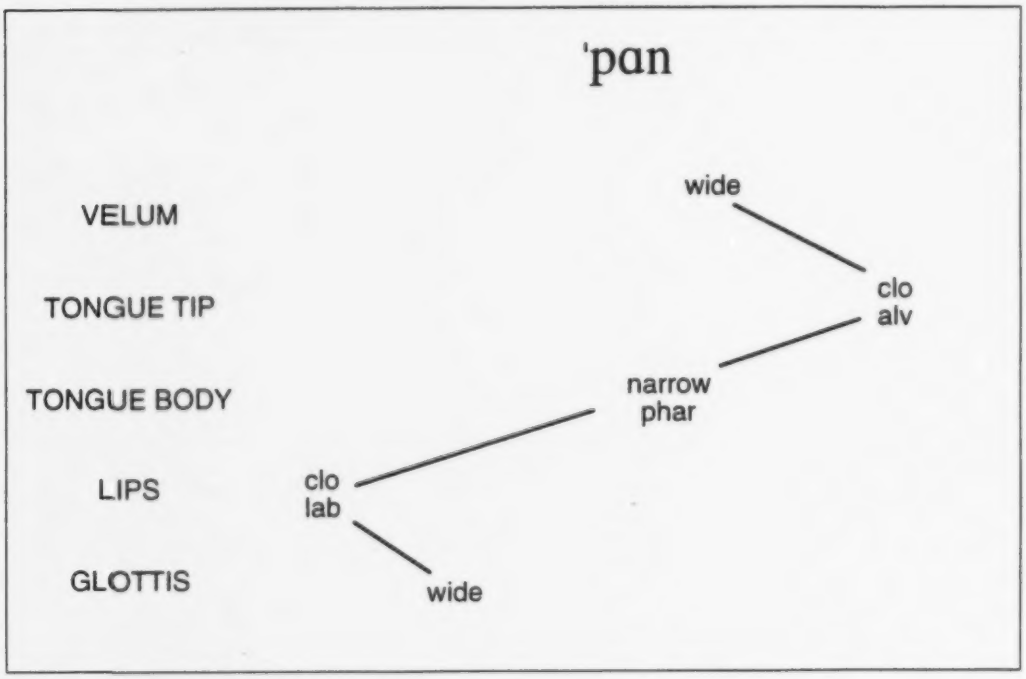
undamped system with the same stiffness as the actual regime. In this way, any characteristic point in the motion of the system can be identified with a phase of this virtual cycle. For example, the movement onset of a gesture is at phase 0 degrees, while the achievement of the constriction goal (the point at which the critically damped system gets sufficiently close to the equilibrium position) occurs at phase 240 degrees. Pairs of gestures are coordinated by specifying the phases of the two gestures that are synchronous. For example, two gestures could be phased so that their movement onsets are synchronous (0 degrees phased to 0 degrees), or so that the movement onset of one is phased to the goal achievement of another (0 degrees phased to 240 degrees), etc. Generalizations that characterize some phase relations in the gestural constellations of English words are proposed in Browman and Goldstein (1990c). As is the case for the values of the dynamic parameters, values of the synchronized phases also appear to cluster in narrow ranges, with onset of movement (0 degrees) and achievement of goal (240 degrees) being the most common (Browman & Goldstein, 1990a).

An example of a gestural constellation (for the word "pawn" as pronounced with the back unrounded vowel characteristic of much of the U.S.) is shown in Figure 3a, which gives an idea of the kind of information contained in the gestural dictionary. Each row, or tier, shows the gestures that control the distinct articulator sets: velum, tongue tip, tongue body, lips, and glottis. The gestures are represented here by descriptors, each of which stands for a numerical equilibrium position value assigned to a tract variable. In the case of the oral gestures, there are two descriptors, one for each of the paired tract variables. For example, for the tongue tip gesture labelled {clo alv}, {clo} stands for -3.5 mm (negative value indicates compression of the surfaces), and {alv} stands for 56 degrees (where 90 degrees is vertical and would correspond to a midpalatal constriction). The association lines connect gestures that are phased with respect to one another. For example, the tongue tip {clo alv} gesture and the velum {wide} gesture (for nasalization) are phased such that the point indicating 0 degrees—onset of movement—of the tongue tip closure gesture is synchronized with the point indicating 240 degrees—achievement of goal—of the velic gesture.

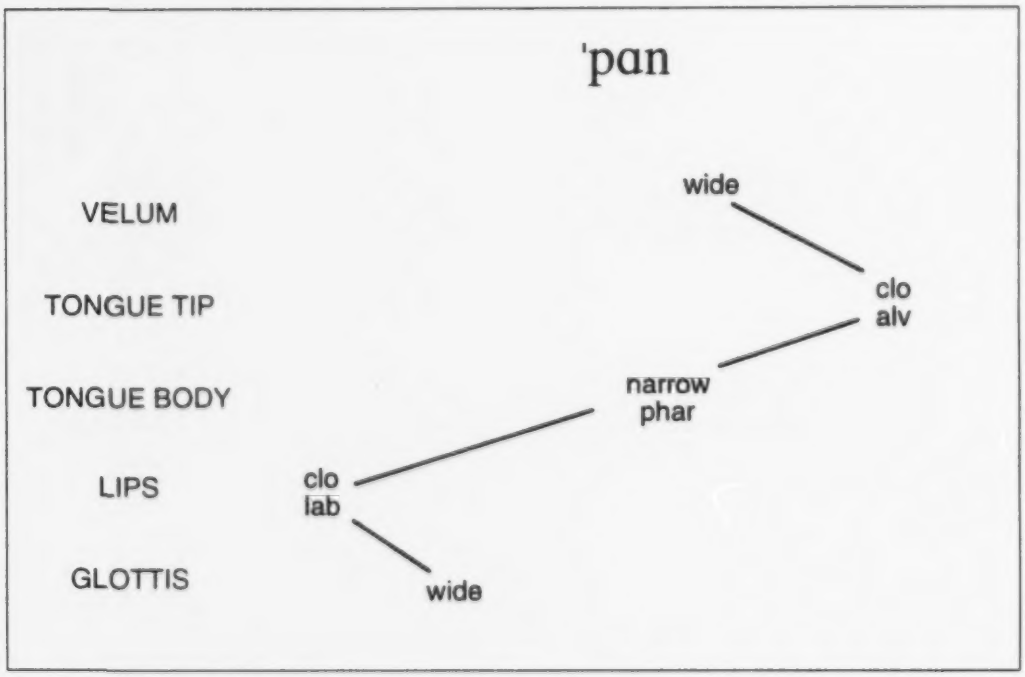
Each gesture is assumed to be active for a fixed proportion of its virtual cycle (the proportion is different for consonant and vowel gestures). The linguistic gestural model uses this proportion, along with the stiffness of each gesture and the phase relations among the gestures, to calculate a *gestural score* that specifies the temporal activation intervals for each gesture in an utterance. One form of this gestural score for "pawn" is shown in Figure 3b, with the horizontal extent of each box indicating its activation interval, and the lines between boxes indicating which gesture is phased with respect to which other gesture(s), as before. Note that there is substantial overlap among the gestures. This kind of overlap can result in certain types of context dependence in the articulatory trajectories of the invariantly specified gestures. In addition, overlap can cause the kinds of acoustic variation that have been traditionally described as allophonic variation. For example in this case, note the substantial overlap between the velic lowering gesture (velum {wide}) and the gesture for the vowel (tongue body {narrow phar}). This will result in an interval of time during which the velo-pharyngeal port is open and the vocal tract is in position for the vowel—that is, a nasalized vowel. Traditionally, the fact of nasalization has been represented by a rule that changes an oral vowel into a nasalized one before a (final) nasal consonant. But viewed in terms of gestural constellations, this nasalization is just the lawful consequence of how the individual gestures are coordinated. The vowel gesture itself hasn't changed in any way: it has the same specification in this word and in the word "pawed" (which is not nasalized).

The parameter value specifications and activation intervals from the gestural score are input to the task dynamic model (Figure 1), which calculates the time-varying response of the tract variables and component articulators to the imposition of the dynamical regimes defined by the gestural score. Some of the time-varying responses are shown in Figure 3c, along with the same boxes indicating the activation intervals for the gestures. Note that the movement curves change over time even when a tract variable is not under the active control of some gesture. Such motion can be seen, for example, in the LIPS panel, after the end of the box for the lip closure gesture. This motion results from one or both of two sources. (1) When an articulator is not part of *any* active gesture, the articulator returns to a neutral position. In the example, the upper lip and the lower lip articulators both are returning to a neutral position after the end of the lip closure gesture. (2) One of the articulators linked to the inactive tract variable may also be linked to some active tract variable, and thus cause passive changes in the inactive tract variable.

(3a)


'pan
VELUM
TONGUE TIP
TONGUE BODY
LIPS
GLOTTIS
wide
clo
alv
narrow
phar
clo
lab
wide

(3b)


'pan
VELUM
TONGUE TIP
TONGUE BODY
LIPS
GLOTTIS
wide
clo
alv
narrow
phar
clo
lab
wide

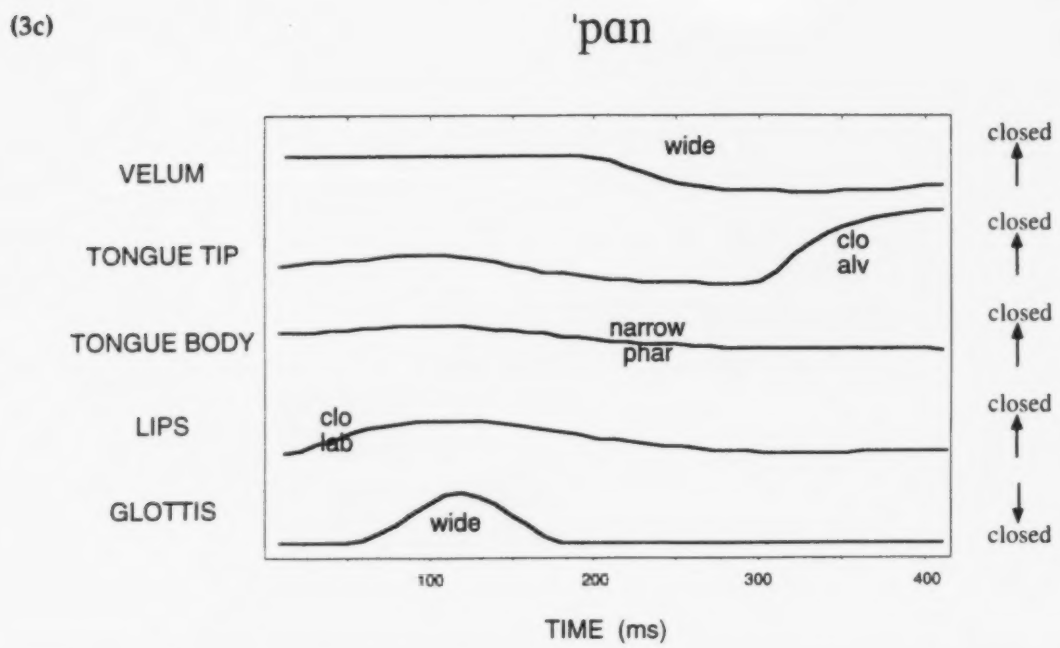


Figure 3. Various displays from computational model for "pawn." (a) Gestural descriptors and association lines (b) Gestural descriptors and association lines plus activation boxes (c) Gestural descriptors and activation boxes plus generated movements of (from top to bottom): Velic Aperture, vertical position of the Tongue Tip (with respect to the fixed palate/teeth), vertical position of the Tongue Body (with respect to the fixed palate/teeth), Lip Aperture, Glottal Aperture.

In the example, the jaw is part of the coordinative structure for the tongue body vowel gesture, as well as part of the coordinative structure for the lip closure gesture. Therefore, even after the lip closure gesture becomes inactive, the jaw is affected by the vowel gesture, and its lowering for the vowel causes the lower lip to also passively lower.

The gestural constellations not only characterize the microscopic properties of the utterances, as discussed above, but systematic differences among the constellations also define the macroscopic property of phonological contrast in a language. Given the nature of gestural constellations, the possible ways in which they may differ from one another is, in fact, quite constrained. In other papers (e.g., Browman & Goldstein, 1986; 1989; 1992) we have begun to show that gestural structures are suitable for characterizing phonological functions such as contrast, and what the relation is between the view of phonological structure implicit in gestural constellations, and that found in other contemporary views of phonology (see also Clements, 1992 for a discussion of these relations). Here we will simply give some examples of how the notion of contrast is defined in a system based on gestures, using the schematic gestural scores in Figure 4.

One way in which constellations may differ is in the presence vs. absence of a gesture. This kind of difference is illustrated by two pairs of subfigures in Figure 4: (4a) vs. (4b) and (4b) vs. (4d). (4a) "pan" differs from (4b) "ban" in having a glottis (wide) gesture (for voicelessness), while (4b) "ban" differs from (4d) "Ann" in having a labial closure gesture (for the initial consonant). Constellations may also differ in the particular tract variable/articulator set controlled by a gesture within the constellation, as illustrated by (4a) "pan" vs. (4c) "tan," which differ in terms of whether it is the lips or tongue tip that perform the initial closure. A further way in which constellations may differ is illustrated by comparing (4e) "sad" to (4f) "shad," in which the value of the constriction location tract variable for the initial tongue tip constriction is the only difference between the two utterances. Finally, two constellations may contain the same gestures and differ simply in how they are coordinated, as can be seen in (4g) "dab" vs. (4h) "bad."

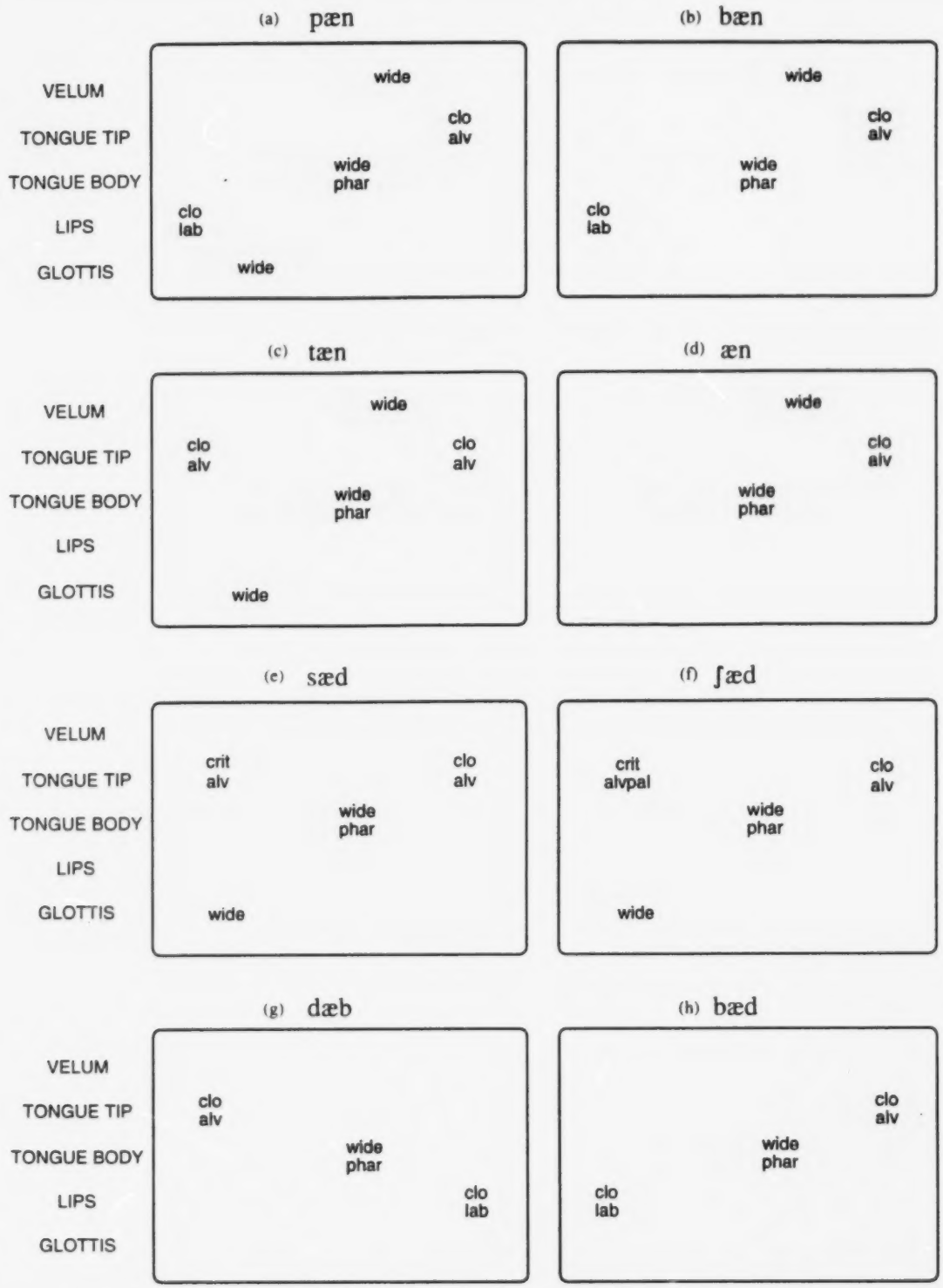


Figure 4. **Schematic gestural scores exemplifying contrast. (a) "pan" (b) "ban" (c) "tan" (d) "Ann" (e) "sad" (f) "shad" (g) "dab" (h) "bad."**

This chapter described an approach to the description of speech in which both the cognitive and physical aspects of speech are captured by viewing speech as a set of actions, or dynamic tasks, that can be described using different dimensionalities: low dimensional or macroscopic for the cognitive, and high dimensional or microscopic for the physical. A computational model that instantiates this approach to speech was briefly outlined. It was argued that this approach to speech, which is based on dynamical description, has several advantages. First, it captures both the phonological (cognitive) and physical regularities that minimally must be captured in any description of speech. Second, it does so in a way that unifies the two descriptions as descriptions with different dimensionality of a single complex system. The latter attribute means that this approach provides a principled view of the reciprocal constraints that the physical and phonological aspects of speech exhibit.

REFERENCES

Abraham, R. H., & Shaw, C. D. (1982). *Dynamics—The geometry of behavior*. Santa Cruz, CA: Aerial Press.

Anderson, S. R., (1974). *The organization of phonology*. New York, NY: Academic Press.

Browman, C. P., & Goldstein, L. (1986). Towards an articulatory phonology. *Phonology Yearbook, 3*, 219-252.

Browman, C. P., & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology, 6*, 201-251.

Browman, C. P., & Goldstein, L. (1990a). Gestural specification using dynamically-defined articulatory structures. *Journal of Phonetics, 18*, 299-320.

Browman, C. P., & Goldstein, L. (1990b). Representation and reality: Physical systems and phonological structure. *Journal of Phonetics, 18*, 411-424.

Browman, C. P., & Goldstein, L. (1990c). In J. Kingston & M. E. Beckman (Eds.), *Tiers in articulatory phonology, with some implications for casual speech* (pp. 341-376).

Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica, 49*, 155-180.

Browman, C. P., Goldstein, L., Kelso, J. A. S., Rubin, P., & Saltzman, E. (1984). Articulatory synthesis from underlying dynamics. *Journal of the Acoustical Society of America, 75*, S22-S23 (A).

Chomsky, N., & Halle, M. 1968. The sound pattern of English. New York: Harper Row. Clements, G. N. (1992). Phonological primes: Features or gestures? *Phonetica, 49*, 181-193.

Clements, G. N. (1992). Phonological primes: Features or gestures? *Phonetica, 49*, 181-193

Cohn, A. C. (1990). Phonetic and phonological rules of nasalization. *UCLA WPP, 76*.

Coleman, J. (1992). The phonetic interpretation of headed phonological structures containing overlapping constituents. *Phonology, 9*, 1-44.

Fourakis, M., & Port, R. (1986). Stop epenthesis in English. *Journal of Phonetics, 14*, 197-221.

Fowler, C. A., Rubin, P., Remez, R. E., & Turvey, M. T. (1980). Implications for speech production of a general theory of action. In B. Butterworth (Ed.), *Language production* (pp. 373-420). New York, NY: Academic Press.

Haken, H. (1977). *Synergetics: An introduction*. Heidelberg Springer-Verlag.

Halle, M. (1982). On distinctive features and their articulatory implementation. *Natural Language and Linguistic Theory, 1*, 91-105.

Hockett, C. (1955). *A manual of phonology*. Chicago: University of Chicago.

Hogeweg, P. (1989). MIRROR beyond MIRROR, Puddles of LIFE. In C. Langton (Ed.), *Artificial life* (pp. 297-316). New York: Addison-Wesley.

Jakobson, R., Fant, C. G. M., & Halle, M. (1951). *Preliminaries to speech analysis: The distinctive features and their correlates*. Cambridge, MA: MIT.

Kauffmann, S. (1989). Principles of adaptation in complex systems. In D. Stein (Ed.), *Sciences of complexity* (pp. 619-711). New York: Addison-Wesley.

Kauffmann, S. (1991). Antichaos and adaptation. *Scientific American, 265*, 78-84.

Kauffmann, S., & Johnsen, S. (1991). Co-evolution to the edge of chaos: coupled fitness landscapes, poised states, and co-evolutionary avalanches. In C. Langton, C. Taylor, J. D. Farmer, & R. Rasmussen (Eds.), *Artificial life II* (pp. 325- 369). New York: Addison-Wesley.

Keating, P. A. (1985). CV phonology, experimental phonetics, and coarticulation. *UCLA WPP, 62*, 1-13.

Keating, P. A. (1988). Underspecification in phonetics. *Phonology, 5*, 275-292.

Keating, P. A. (1990). Phonetic representations in a generative grammar. *Journal of Phonetics, 18*, 321-334.

Kelso, J. A. S., Saltzman, E. L., & Tuller, B. (1986). The dynamical perspective on speech production: Data and theory. *Journal of Phonetics, 14*, 29-59.

Kent, R. D., & Minifie, F. D. (1977). Coarticulation in recent speech production models. *Journal of Phonetics, 5*, 115-133.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America, 59*, 1208-1221.

Kugler, P. N., & Turvey, M. T. (1987). *Information, natural law, and the self- assembly of rhythmic movement*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Ladefoged, P. (1980). What are linguistic sounds made of? *Language, 56*, 485- 502. Ladefoged, P. (1982). *A course in phonetics* (2nd ed.). New York, NY: Harcourt Brace Jovanovich.

Lewin, R. (1992). *Complexity*. New York, NY: Macmillan.

Ladefoged, P. (1982). *A course in phonetics* (2nd ed.). New York: Harcourt Brace Jovanovich.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74*, 431- 461.

Liberman, M., & Pierrehumbert, J. (1984). Intonational invariance under changes in pitch range and length. In M. Aronoff, R. T. Oehrle, F. Kelley, & B. Wilker Stephens (Eds.), *Language sound structure* (pp. 157-233). Cambridge, MA: MIT Press.

Lindblom, B., MacNeilage, P., & Studdert-Kennedy, M. (1983). Self-organizing processes and the explanation of phonological universals. In B. Butterworth, B. Comrie, & O. Dahl (Eds.), *Explanations of linguistic universals* (pp. 181-203). Mouton: The Hague.

Macchi, M. (1988). Labial articulation patterns associated with segmental features and syllable structure in English. *Phonetica 45*, 109-121.

Madore, B. F., & Freedman, W. L. (1987). Self-organizing structures. *American Scientist, 75*, 252-259.

Manuel, S. Y., & Krakow, R. A. (1984). Universal and language particular aspects of vowel-to-vowel coarticulation. Haskins *Laboratories Status Report on Speech Research, SR77/78*, 69-78.

McCarthy, J. J. (1988). Feature geometry and dependency: A review. *Phonetica, 45*, 84-108.

Ohala, J. (1983). The origin of sound patterns in vocal tract constraints. In P. F. MacNeilage (Ed.), *The production of speech* (pp. 189-216). New York, NY: Springer-Verlag.

Öhman, S. E. G. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America, 39*, 151- 168.

Packard, N. (1989). Intrinsic adaptation in a simple model for evolution. In C. Langton (Ed.), *Artificial life* (pp. 141-155). New York: Addison-Wesley.

Pierrehumbert, J. (1990). Phonological and phonetic representation. *Journal of Phonetics 18*, 375-394.

Pierrehumbert, J. B., & Pierrehumbert, R. T. (1990). On attributing grammars to dynamical systems. *Journal of Phonetics, 18*.

Port, R. F. (1981). Linguistic timing factors in combination. *Journal of the Acoustical Society of America, 69*, 262-274.

Rubin, P. E., Baer, T., & Mermelstein, P. (1981). An articulatory synthesizer for perceptual research. *Journal of the Acoustical Society of America, 70*, 321-328.

Sagey, E. C. (1986). *The representation of features and relations in non-linear phonology*, doctoral dissertation, MIT.

Saltzman, E. (in press). Dynamics in coordinate systems in skilled sensorimotor activity. In T. van Gelder & B. Port (Eds.), *Mind as motion*. Cambridge, MA: MIT Press.

Saltzman, E. (1986). Task dynamic coordination of the speech articulators: A preliminary model. In H. Heuer & C. Fromm (Eds.), *Generation and modulation of action patterns* (pp. 129-144). Berlin/Heidelberg: Springer-Verlag

Saltzman, E., & Kelso, J. A. S. (1987). Skilled actions: A task dynamic approach. *Psychological Review, 94*, 84-106.

Saltzman, E. L., & Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology 1*, 333-382.

Schoner, G., & Kelso, J. A. S. (1988). Dynamic pattern generation in behavioral and neural systems. *Science, 239*, 1513-1520.

Stevens, K. N. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. In E. E. David & P. B. Denes (Eds.), *Human communication: A unified view* (pp. 51-66). New York, NY: McGraw-Hill.

Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics, 17*, 3-45.

Sussman, H. M., MacNeilage, P. F., & Hanson, R. J. (1973). Labial and mandibular dynamics during the production of bilabial consonants: Preliminary observations. *Journal of Speech and Hearing Research, 16*, 397-420.

Turvey, M. T. (1977). Preliminaries to a theory of action with reference to vision. In R. Shaw & J. Bransford (Eds.), *Perceiving, acting and knowing: Toward an ecological psychology*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Wood, S. (1982). X-ray and model studies of vowel articulation. *Working Papers*, Lund University, *23*.

FOOTNOTES

*In T. van Gelder & B. Port (Eds.), *Mind as motion*. Cambridge, MA: MIT Press (in press).

† Also Department of Linguistics, Yale University.

Haskins Laboratories Status Report on Speech Research
1993, SR-113, 63-90

Some Organizational Characteristics of Speech Movement Control*

Vincent L. Gracco

The neuromotor organization for a class of speech sounds (bilabials) was examined to evaluate the control principles underlying speech as a sensorimotor process. Oral opening and closing actions for the consonants /p/, /b/, and /m/ (C1) in /s V1 C1 V2 C2/ context, where V1 was either /ae/ or /i/, V2 was /ae/, and C2 was /p/, were analyzed from four subjects. The timing of oral opening and closing action was found to be a significant variable differentiating bilabial consonants. Additionally, opening and closing actions were found to covary along a number of dimensions implicating the movement cycle as the minimal unit of speech motor programming. The sequential adjustments of the lips and jaw varied systematically with phonetic context reflecting the different functional roles of these articulators in the production of consonants and vowels. The implication of these findings for speech production is discussed.

INTRODUCTION

As a motor process, speaking is an intricate orchestration of multiple effectors (articulators) coordinated in time and space to produce sound sequences. From a functional perspective, the speech production mechanism is a special purpose device modulating the aerodynamic and resonance properties of the vocal tract by rapidly creating constrictions, occlusions, or overall shape changes. These events provide the foundation for the categorically distinct sounds of the language. The functional task, driven by cognitive considerations, involves a number of sensorimotor processes that generate segmental units, specify the coordination among contributing articulators, scale articulator actions to phonetic and pragmatic contexts, and subsequently sequence these actions into meaningful units for communication.

This research was supported by grants DC-00121, and DC-00594, from the National Institute of Deafness and Other Communication Disorders The author thanks Carol Fowler, Anders Löfqvist, Ignatius Mattingly and Rudolph Sock for comments on earlier versions of this manuscript. The author also thanks John Folkins for a thorough and constructively critical review of this manuscript.

Assuming that the individual phonetic segments of a language are stored in some manner in the nervous system, a number of important issues can be identified. What are the production units, how are they differentiated for sounds, and how are they modified according to context? The present investigation is an attempt to answer some of these questions and identify some principles of speech motor organization used to scale, coordinate, and sequence multiple articulatory actions.

Movement characteristics—Stop consonants

One focus of the present investigation is to examine, in detail, the kinematic adjustments of lip and jaw motion associated with a class of English speech sounds, the bilabial stop consonants /p/, /b/, /m/, in different vowel contexts. These consonants are the same in their place of articulation (the lips), employing the same set of articulators to create an obstruction at the oral end of the vocal tract. One property distinguishing these sounds is the presence or absence of voicing; /p/ is a voiceless consonant because laryngeal vibration is briefly (100-200 msec) arrested during the oral closing. As the lips release the oral closure the vocal folds move back together and voicing for the next vowel is initiated (Lisker &

Abramson, 1964). There are two other classes of stop consonants in English which create temporary obstructions in different regions of the vocal tract, and use different articulators; alveolar and velar stop consonants. In addition, each of the alveolar and velar voiced consonants can be produced with the velum lowered, creating the nasal resonances for the /n/ and /ŋ/ sounds, respectively. Thus, English contains three characteristic sets of stop consonants that differ in the primary articulators used to create the obstruction and the location of the obstruction in the vocal tract.

It is not clear from previous investigations how articulatory movements within a class are modified by concomitant articulatory actions. That is, are the lip and jaw movements for all bilabial consonants similar with the acoustic distinctions between sounds in the class attributed solely to the laryngeal and velar actions? Examination of electromyographic (EMG), kinematic, and acoustic data reveal conflicting findings. Electromyographic activity from lip muscles for bilabial production has generally failed to reveal significant differences in measured variables such as peak EMG amplitude and EMG burst duration associated with the oral closing action (Fromkin, 1966; Harris, Lysaught, & Schvey, 1965; Lubker & Parris, 1970; Tatham & Morton, 1968). The one exception is the study by Sussman, MacNeilage, & Hanson (1973) in which the activity from one upper lip depressor muscle (depressor anguli oris) from one subject was found to be greater for /p/ than for /b/ and /m/. Kinematic studies have revealed a few movement differences for /p/ and /b/, most consistently a tendency for the oral closing movement velocity to be higher for /p/ than /b/ or /m/ (Chen, 1970; Summers, 1987; Sussman et al., 1973). Two of the above mentioned investigations (Sussman et al., 1973; Summers, 1987) have also reported kinematic and EMG differences in movements surrounding the oral closing action. Jaw opening for the vowel /a/ or /ae/ before /p/ was found to be higher than before /b/ (Summers, 1987) and the lower lip opening velocity and associated EMG activity (following the oral occlusion) was greater for /m/ than /p/ or /b/ (Sussman et al., 1973).

Acoustic studies have provided the most reliable findings associated with voicing differences. Closure durations for voiceless consonants are generally longer than closure durations for their voiced cognates and the preceding vowel is shorter in the voiceless context (Denes, 1955; House & Fairbanks, 1953; Luce & Charles-Luce, 1985). Integrating results from different empirical observations allows for some potential speculation on the voiced/voiceless differences. EMG results suggest that the form of the oral closing motor commands is not different in magnitude or duration due to the presence or absence of voicing. In contrast, the movement and acoustic studies suggest that differences may be focused on movement timing as evidenced by changes in movement velocity and vowel duration. While the available data suggest that one difference between sounds within an equivalence class such as bilabials may be reflected in aspects of their timing, no detailed analysis has been conducted.

Another possibility is that articulatory motion may be governed by control principles that operate on combinations of kinematic variables reflecting important system parameters. For example, consonant related movement differences may be specified by articulator stiffness, a construct with some potential as a motor control variable (Cooke, 1980; Ostry & Munhall, 1985; Kelso et al., 1985). If an effector such as the lower lip is modeled as a linear second order system, changing the static stiffness of the system results in predictable changes in the velocity/displacement relationship; the velocity/displacement ratio varies directly with stiffness. Moreover if resting stiffness is a variable being controlled by the nervous system then changes in an effector's kinematics can be brought about by a single system parameter. It has been suggested that mass normalized stiffness, estimated from the ratio of a movement's peak velocity and displacement, may be an important control parameter for skilled actions including speech (Kelso et al., 1985; Kelso, 1986). While a number of studies have demonstrated consistent and systematic velocity/displacement relations for a range of speech movements (Kuehn & Moll, 1976; Munhall, Ostry & Parush, 1985l; Ostry, Keller & Parush, 1983), to date it has not been demonstrated unequivocally that stiffness is an important control parameter, except as one of many possible alternatives (see Nelson, 1983 for example). One possibility is that /p/ with higher closing velocity might be differentiated from /b/ and /m/ by its stiffness specification suggesting that speech sounds may be coded neurophysiologically by such a construct.

Articulator interactions and sequencing

It is becoming increasingly clear that individual articulators and their respective actions are not independent but functionally related by task. The empirical support for this view comes from two

main sources. When the motions of the lips and jaw for bilabial closure are examined in detail, it appears that the individual articulators are not adjusted independently either in magnitude or relative timing (Gracco, 1988; Gracco & Abbs, 1986; Hughes & Abbs, 1976). Spatiotemporal adjustments to suprasegmental manipulations also suggest that stress and rate mechanisms act on all active regions of the vocal tract (Fowler, Gracco, V.-Bateson, & Romero, submitted). Changes in articulator movement patterns following mechanical perturbation are also consistent with the suggestion that the actions of individual articulators are not adjusted independently. Rather, disruptions to articulator movement during speech result in adjustments in the perturbed articulator as well as those (unperturbed) articulators that are actively involved in the production of the specific sound (Folkins & Abbs, 1975; Kelso et al., 1984; Kollia, Gracco, & Harris, 1992; Munhall, Löfqvist, & Kelso, in press; Shaiman, 1989; Abbs & Gracco, 1984; Gracco & Abbs, 1985; 1988). While many previous studies can be criticized on statistical grounds (see Folkins & Brown, 1987) it is not unreasonable to suggest that interactions among articulators are not random but systematic and integral to the overall speech production process. Together these observations suggest that speech movements are organized according to higher level principles that involve articulatory aggregates (similar to the principle of synergy defined by Bernstein, 1967).

Speaking is also a continuous process involving articulatory motion generating acoustic and aerodynamic events. Lip and jaw motion can be classified as either opening or closing depending on the direction of the motion relative to an occlusion/constriction or characteristic vocal tract shape. Oral opening is most often associated with vowel production while oral closing is most often associated with consonant production. Different relative timing patterns for oral articulatory motions depending on direction have been interpreted as manifestations of separate and distinct synergistic actions fundamental to the speech production process (Gracco, 1988). Moreover, the opening phase of an articulatory action shows greater variation in spatiotemporal adjustment due to stress (Kelso, V.-Bateson, Saltzman, & Kay, 1985; Ostry, Keller, & Parush, 1983) than the corresponding closing phase. Similarly, oral opening and closing movement duration show differential effects to changes in speaking rate (Adams, Weismer, & Kent, 1993). These differential patterns associated with movement direction are also influenced by phonetic context. The context in which a sound is produced is known to affect the acoustic and kinematic properties associated with its production (Daniloff & Moll, 1968; Sussman, MacNeilage, & Hanson, 1973; Parush, Ostry, & Munhall, 1983; Perkell, 1969). It seems that articulatory adjustments for context may have differential effects on the different movement phases. In order to understand the organizational principles and control processes that govern speech production, evaluation of speech movement differences must include the articulatory adjustments within as well as across articulators and movement phases.

Examining kinematic changes across movement phases has additional implications for understanding speech as a serial process. As pointed out by Lashley (1951), serial actions, such as those found in speech, locomotion, typing, and the playing of musical instruments, cannot be explained in terms of successions of external stimuli or reflex chaining. Rather, the apparent rhythmicity found in all but the simplest motor activities suggests that some sort of temporal patterning or temporal integration may form the foundation for motor as well as perceptual activities (Lashley, 1951). Lashley further suggested that skilled action involves the advanced planning of entire sequences of action. More recently, Sternberg and colleagues have developed a model of speech timing that incorporates the concept of an advanced plan of action, an utterance program, which is used to control the execution of the elements in sequence (Sternberg, Knoll, Monsell, & Wright, 1988). The elements of the utterance program are action units defined abstractly on the basis of requiring a single selection process but may contain multiple distinguishable actions (Sternberg et al., 1988). That is, speech production is considered a hierarchical process with advanced planning of what is to be said (utterance program) followed by the execution of the program using smaller sequenced elements (action units) on the order of stress groups (Fowler, 1983; Sternberg et al., 1988).

While this model provides a framework for the speech motor process, a number of unresolved issues remain. Previous work has focused on evaluating the latency and duration of words or syllables rapidly produced by subjects. Examination of articulatory movement has not been undertaken and thus identification of the articulatory correlates of the action unit is lacking. If an action unit or unit of speech production is to have any theoretical significance, identification of the physiological (articulatory)

instantiation of the construct is crucial. Further, in order to fully understand the speech production process, the mechanism by which elemental units are sequenced and modified is also of central import. In the present investigation, serial speech movements within and across movement sequences were examined for evidence of articulatory cohesiveness that may reflect the production unit as well as the manner in which phonetic context may modulate the underlying temporal patterning.

Articulatory coordination

A final focus of the present investigation is the coordination patterns of the lips and jaw as they cooperate to occlude and open the oral end of the vocal tract for the different bilabial consonants. A number of previous investigations suggest that task-related articulatory movements, such as the lip and jaw motion for bilabial closing, are interdependent in their timing (Gracco, 1988; Gracco & Abbs, 1986). Results presented by Löfqvist and Yoshioka (1981;1984) similarly suggest a number of consistent timing patterns for different laryngeal-oral interactions following stress, rate, and consonantal manipulations. Observations of timing coherence among articulatory actions have been extended to include the lip, jaw, and larynx associated with the initiation of voicing and devoicing in a variety of phonetic contexts (Gracco, 1990; Gracco & Löfqvist, 1989; Gracco & Löfqvist, in preparation). Such observations reflect a potential principle of speech movement organization. For speech movement coordination, as for motor coordination in general (see Bernstein, 1967), the timing or patterning of the potential degrees of freedom for task-related actions, are constrained thereby reducing the overall motor control complexity (Gracco, 1988).

While the available evidence suggests that constraining the degrees of freedom may be a general principle underlying the coordination of task-related speech articulators, it has recently been suggested that lip and jaw coordination is not invariant (DeNil & Abbs, 1991). Rather, variations in the temporal sequencing of lip and jaw closing movements for /b/ at fast and slow speaking rates, have been interpreted to reflect fundamentally different patterns requiring different coordinative patterns. With the exception of the DeNil and Abbs (1991) investigation, previous studies examining the relative timing of the lips and jaw during oral closing have focused on voiceless /p/ (Caruso, Abbs, & Gracco, 1988; Gracco & Abbs, 1986; 1988; Gracco, 1988; McClean, Kroll, & Loftus, 1990). It is possible that lip and jaw actions for voiced and voiceless sounds differ in their relative timing, reflecting different coordinative relations. Further, one previous study has examined the relative timing of the lips and jaw for oral opening, and reported fundamentally different timing patterns among the articulators for oral opening than for oral closing (Gracco, 1988). As such, extended examination of interarticulatory timing across movement phases may be informative regarding the extent and generality of any coordinative principles governing speech motor actions.

Methods

Subjects and movement task. Four females between the ages of 21 and 28 years were subjects in the present study. None had a history of neurological disorder and all were native speakers of American English. Subjects were asked to repeat one of six utterances following the onset of an experimenter-controlled tone. The utterances were of the form /s V1 C V2/ where V1 was either /ae/ or /i/, C was /p/, /b/, or /m/ and V2 was /ae/. The specific utterances were:

1) sapapple
2) seepapple
3) sabapple
4) seebapple
5) samapple
6) seemapple

Each word was repeated, in isolation, a total of 40 times, 10 or 20 consecutive times for each word on the list in the presented order, followed by subsequent repetition(s) of the entire list. A total of 240 tokens were obtained for each subject with the exception of subject four whose total was 238 (one sapapple and one seemapple missing). The stimulus to respond (auditory tone) was presented at approximately three second intervals. Subjects were instructed to produce each word with equal stress at a comfortable speaking rate following the tone and to use an effort level appropriate for speaking to someone 10-15 feet away.

Instrumentation and data acquisition. Single dimensional midsagittal movements of the upper lip (UL), lower lip (LL), and jaw (J) were transduced using a head-mounted strain gage system previously described (Barlow, Cole, & Abbs, 1983). Briefly, movements of the upper lip, lower lip, and jaw were transduced using ultralight weight cantilever beams instrumented with strain gages attached to a lightweight head-mounted frame. Transducers were attached midsagittally at the vermilion border of the upper and lower lips and on a region of the chin which

yielded minimal skin movement artifact (see Barlow et al., 1983; Folkins, 1981; Kuehn, Reich, & Jordan, 1980; Müller & Abbs, 1979). Transducers were visually aligned to capture the major movement axis for oral opening/closing for each subject with the resultant orientation rotated approximately 100 to 110 degrees posterior to the vertical. Movement signals were digitized at a sampling rate of 500 Hz with 12 bit resolution except for subject four in which movement signals were sampled at 400 Hz. The acoustic signal was sampled at 500 Hz to indicate the onset and offset of the stimuli only.

Data analysis. Prior to data analysis, all movement signals were digitally low pass filtered (Butterworth implementation, two-pole zero phase lag, 20 Hz cutoff). As transduced, the lower lip signal reflects the combined movement of the lower lip and jaw. In order to obtain lower lip movement only, the jaw signal was software subtracted from the lower lip signal, yielding net lower lip movement. First derivatives were obtained using a three-point numerical differentiation routine (central difference) and stored as part of the data file. From these instantaneous velocity signals, movement onsets and offsets were determined for the individual closing movements. Computer software was used to identify zero crossings in the velocity traces of the respective upper lip, lower lip, and jaw signals and movement onset and offsets were marked. In addition, the movements of the upper lip, lower lip, and jaw were combined to yield a measure of the overall lip aperture (LA).

For the present experiment, subjects had been instructed to start each utterance from a position of rest with the lips in contact, and the upper and lower teeth lightly touching. Using this relatively consistent starting point it was possible to evaluate the relative position of the different articulators for the /s/ and the following vowel.

Each acquired token consisted of eight signals illustrated in Figure 1. Two complete movement cycles or sequences were identified from the lip aperture signal (LA) and indicated as sequence 1 and 2 (S_1, S_2) in the figure. Each sequence consists of two phases, an opening and closing phase. Sequences were defined as the time interval between the onset of the opening movement to the offset of the closing movement in the LA signal using a zero velocity criteria. Onsets of the opening and closing movements for the lips and jaw were also determined from zero crossings in the first derivatives of the respective articulator motions. Maximum displacement and duration were defined as the distance and time, respectively, between the onset and offset of each movement; peak instantaneous velocity was also obtained for each movement. As with previous studies (Gracco, 1988; Gracco & Abbs, 1986) a measure of upper lip, lower lip, and jaw coordination during oral closing was obtained from the timing of each articulator's peak velocity relative to the same preceding articulatory event (jaw opening for the vowel). From the identified events of interest, the following measures were made (see Figure 1):

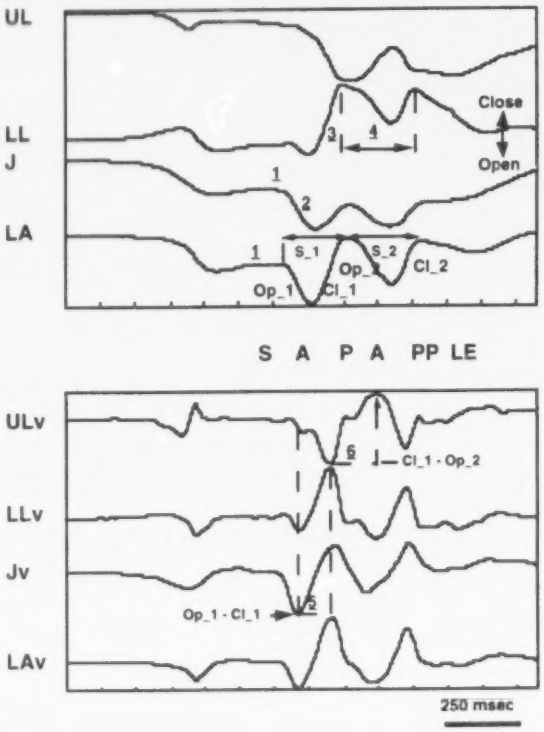


Figure 1. A representative example of upper lip (UL), lower lip (LL), jaw (J), and lip aperture (LA) displacement (above) and velocity (below) signals for one of the stimulus words, "sapapple." All trials were initiated from a rest position with the lips together and teeth lightly touching. Each production included two opening/closing sequences identified from the LA signal as S_1 and S_2. Within each sequence or cycle, opening and closing phases were also identified (Op_1, Cl_1, Op_2, Cl_2). Measurements are indicated by the numbers in the figure (see text for details).

1) extent of oral opening and associated jaw position for /s/, relative to the subjects' rest position,

2 extent of oral opening for the first vowel (Op_1) and the associated jaw opening movement characteristics,

3) lip and jaw closing movement characteristics for the first closing,

4) oral opening (Op_2) and closing (Cl_2) movement characteristics for the second cycle (S_2) for the /aep/ in "apple,"

5) time of upper lip, lower lip, and jaw peak closing velocity relative to the peak jaw *opening* velocity for the first vowel opening,

6) time of upper lip, lower lip, and jaw peak opening velocity for the second opening relative to the peak upper lip *closing* velocity for the first oral closing.

Statistical analysis. The data were analyzed using two factor repeated measures ANOVA with vowels and consonants as factors. When there were significant interactions, post hoc analyses were conducted using Scheffe's F-test with alpha level set to .01 for all comparisons.

Results

Context-dependent changes in the opening and closing movements will be examined for the first movement sequence. Next, movement parameters of the second sequence will be examined for carryover effects from the phonetic context of the first sequence. Finally, the coordination of lip and jaw motion will be examined for closing and opening effects and phonetic context. Individual articulator adjustments will be concurrently examined to evaluate their functional role in the speech production process.

Movement characteristics

Sequence One

Cycle duration. One of the most robust findings in the present study was a change in the duration of the first movement cycle. For the group, the cycle duration obtained from the lip aperture signal was, on average, shorter for the /i C/ than /ae C/ context (191.8 msec Vs 227.3 msec). In addition, the cycle duration was shorter when the consonant was /p/ than /b/ or /m/ (203.3 vs. 210.2 vs. 215.1 msec, respectively). Shown in Figure 2 are three examples of the lower lip and jaw movements for "sapapple," "sabapple" and "samapple" from a single subject. All movements in the figure were aligned to the same articulatory event; jaw opening peak velocity for the first vowel /ae/ (dotted line). As illustrated by the arrows in Figure 2, the interval between the J opening peak velocity for /ae/ and the time of maximum lip and jaw closing velocity and displacement is shortest when the consonant is /p/. Because the jaw opening peak velocities are aligned in these examples, it can also be seen that the adjustment within the cycle was localized to the interval spanning the terminal phase of the opening action and the time of the peak velocity for the closing action. This interval for all phonetic contexts and subjects for the first movement cycle is presented in Figure 3. Since there were no articulator specific differences, only the LL will be considered.

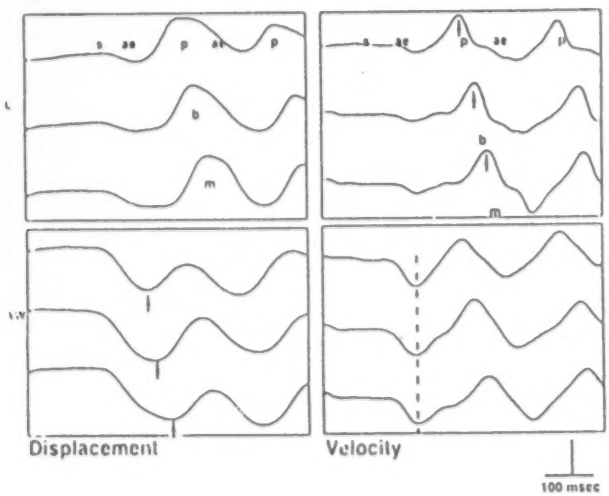


Figure 2. Three representative lower lip (LL) and jaw movements (displacement and velocity) from S2 for "sapapple," "sabapple," and "samapple." All signals were aligned to the jaw opening peak velocity (dotted line). Opening is toward the bottom; closing is toward the top). Arrows in jaw opening displacement panel (lower left) indicate the shortening of the jaw opening movement duration for the different consonants. Similarly, the LL closing movement for /p/ achieves both peak velocity (arrows, upper right panel) and peak displacement (upper left panel) earlier for /p/ than /b/ and /m/. Vertical calibration is 6 mm (displacement) and 150 mm/sec (velocity).

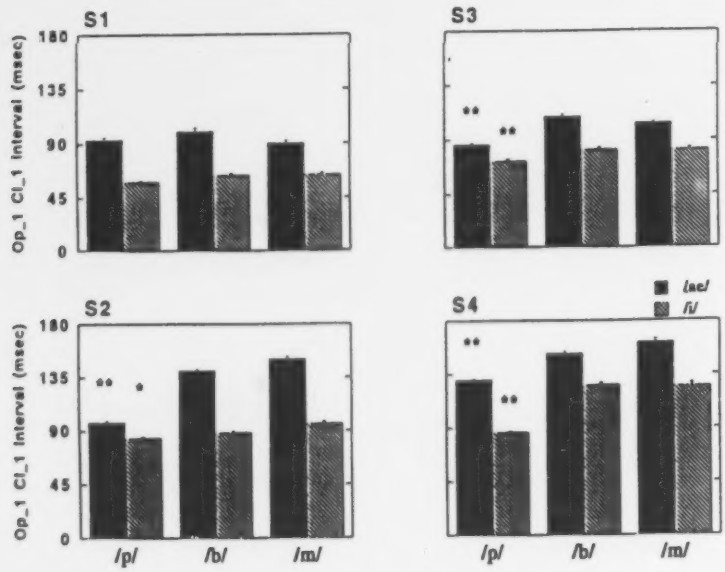


Figure 3. **Average interval (in milliseconds) between the jaw opening peak velocity for the first vowel (Op_1) and the time of the LL peak velocity for the first oral closing (Cl_1) for the different phonetic contexts for all subjects. Single asterisk (*) reflects a significant /p/ - /b/ or /p/ - /m/ difference; a double asterisk (**) indicates that both comparisons were reliable ($p < .01$).**

Table 1. *Mean difference (x_d) and associated F-value (Scheffe's F-test) for within-vowel (/aep-aeb/, /aep-aem/, /ip-ib/, /ip-im/) consonant comparisons and across-vowel (/aep-ip/, /aeb-ib/, /aem-im/) consonant comparisons for the Op_1 Cl_1 interval and lower lip closing movement duration for all subjects. Degrees of freedom for S1-3 (5, 234), S4 (5,232); asterisk indicates a significant difference at $p < .01$.*

Op_1 Cl-1 interval

	/aep-aeb/	/aep-aem/	/ip-ib/	/ip-im/	/aep-ip/	/aeb-ib/	/aem-im/
S1 (x_d)	-6.75	3.25	-5.4	-5.75	35.55	36.9	26.55
F	1.15	.27	.73	.83	31.8*	34.26*	17.73*
S2	-43.5	-53.55	-4.4	-12.1	13.5	52.6	54.95
	75.99*	115.16*	.78	5.68*	7.32*	111.11*	121.26*
S3	-22.95	-17.5	-10.05	-9.75	14.6	27.5	22.35
	18.9*	10.99*	3.62*	3.41*	7.65*	27.14*	17.93*
S4	-22.1	-31.29	-39.69	-38.99	44.71	27.12	37.01
	12.88*	25.81*	42.06*	40.09*	52.71*	19.65*	36.11*

Movement duration

	/aep-aeb/	/aep-aem/	/ip-ib/	/ip-im/	/aep-ip/	/aeb-ib/	/aem-im/
S1 (x_d)	-5.05	-3.35	.3	-5.1	-2.8	2.55	-4.55
F	1.93	.85	.01	1.97	.59	.49	1.57
S2	-19.2	-11.8	-.15	-13.2	-18.25	.8	-19.65
	11.74*	4.44*	.00	5.55*	10.61*	.02	12.3*
S3	-20.2	-22.45	-3.65	-6.35	-13.1	3.45	3
	9.98*	12.32*	.33	.99	4.2*	.29	.22
S4	-4.66	-5.16	-.94	-15.38	-3.6	.12	-13.81
	.51	.63	.02	5.59*	.31	.0003	4.51*

Within vowel and consonant comparisons revealed a significantly shorter interval when the consonant was /p/ than /b/ and /m/ in both vowel contexts, and a longer interval when the vowel was /ae/ compared to /i/. These effects were reliable for all subjects except for S1 whose consonant effect did not reach significance. Inspection of the data from S1 indicated that for the first block of ten repetitions the same trend as seen in the other subjects was present. However, for the subsequent blocks, the overall speech rate was increased resulting in a reduction in the /p/ - /b/ - /m/ interval differences. In contrast to the interval results, the LL closing movement durations were not consistently affected by the consonant or vowel identity. Table 1 presents the average Op_1 Cl_1 interval and movement duration differences for the LL for each subject and the associated F statistic (see Table 1).

Oral opening--Position/Displacement/Velocity

Since the subjects began each experimental trial with their lips together and teeth lightly touching, the opening position for the first sound (/s/) in the movement sequence could be examined for anticipatory adjustments associated with the different phonetic context (upcoming vowel and/or consonant). The vertical oral opening for /s/ obtained from the lip aperture signal, averaged 6.78 mm (S1) to 9.38 mm (S2) across subjects and conditions and was not affected by subsequent phonetic context ($p > .2$). In contrast, the maximum oral opening for the vowel was highly dependent on vowel identity. Figure 4 presents the average oral opening for the group and the individual lip and jaw contributions to the opening for the different vowel-consonant contexts. The oral opening was significantly larger for /ae/ compared to /i/ for all subjects (S1[$F(1,238) = 230.81$, $p = .0001$]; S2[$F(1,238) = 2188.6$, $p = .0001$]; S3[$F(1,238) = 796.51$, $p = .0001$]; S4[$F(1,236) = 343.96$, $p = .0001$]). Oral opening for /ae/ averaged 15.2, 17.6, 17.9, and 13.3 mm, while oral opening for /i/ averaged 12.2, 9.8, 11.5, and 10.0 mm for S1-4 respectively. As can be seen from the consistent UL plus LL contribution to the oral opening in Figure 4, the jaw is the articulator solely responsible for the oral opening changes. Jaw position was consistently higher for /i/ than /ae/ for all subjects (S1[$F(1,238) = 535.41$, $p = .0001$]; S2[$F(1,238) = 2067.8$, $p = .0001$]; S3[$F(1,238) = 610.89$, $p = .0001$]; S4[$F(1,236) = 505.12$, $p = .0001$]). Collapsed across consonants jaw opening position averaged 3.4, 7.0, 6.1, and 3.5 mm higher for /i/ than /ae/ (S1-4 respectively). The voicing character of the consonant had no consistent effect on the jaw opening position. No significant vowel or consonant effects were noted for the UL+LL.

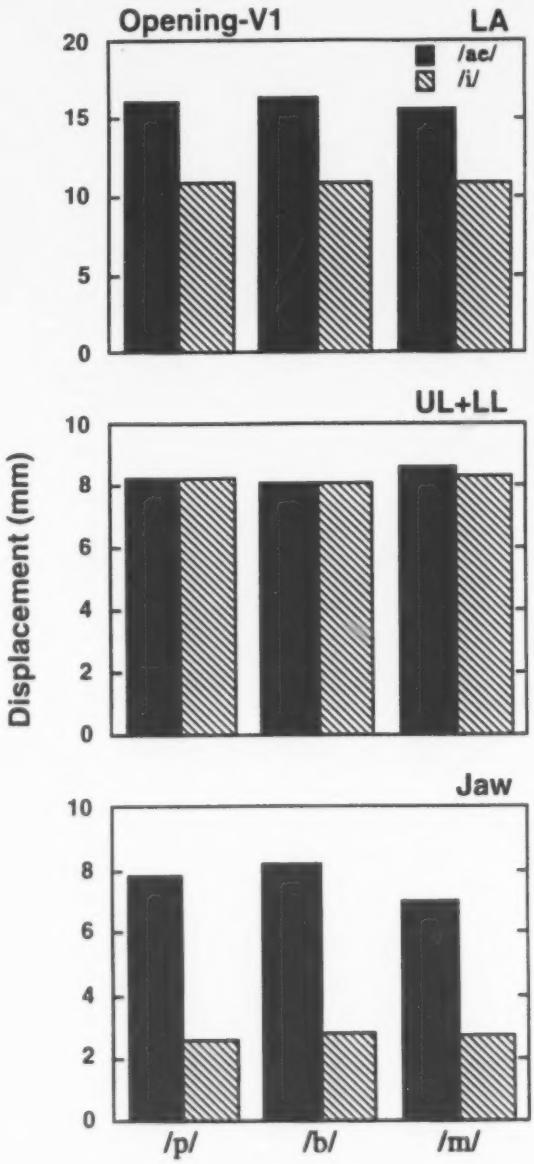


Figure 4. **Group lip aperture (LA) opening, in millimeters (mm) relative to the rest position, for the first vowel (V1) and consonant contexts. Below the LA data are the contributions of the UL plus LL (UL+LL) and jaw to the oral opening Vertical bars indicate one standard deviation.**

In the present phonetic context, it was often not possible to unambiguously identify UL or LL opening movement associated with the vowel from movements associated with /s/. Thus, the opening velocity for the first movement sequence focused exclusively on the jaw. Because of the large vowel related jaw opening displacement effect, jaw opening characteristics were examined for consonant-related effects only. Figure 5 presents the jaw opening velocity results for the four subjects. Post hoc testing revealed significant consonant effects in the /ae/ context only with jaw opening velocity higher before /p/ than /m/ for all subjects (S1[$F(2,117) = 8.11$, $p < .01$]; S2 $F = 30.97$, $p < .01$]; S3 $F = 29.88$, $p < .01$]; S4[$F(2,116) = 9.0$, $p < .01$]); /p/ - /b/ differences were only significant for S2 ($F = 9.0$, $p < .01$)

Oral closing—Displacement

As a consequence of changes in oral opening displacement for the different vowels, oral closing movements were modified in their extent. As shown in Figure 6 for the group, movement extent for each articulator is greater for the more open vowel /ae/. Presented in Table 2 are the individual subject comparisons for the upper lip, lower lip, and jaw closing displacements. There was a highly significant vowel effect for the J for all subjects with larger closing movement displacements in the /ae/ context compared to /i/. Results for the UL were in general agreement with those of the J but to a lesser degree; S2, S3, and S4 demonstrated significantly larger closing displacements in the /ae/ context compared to /i/. The LL results were less consistent with fewer comparisons reaching significance. No significant UL or LL differences related to vowel identity were noted for S1. Consonant related movement adjustments were much less consistent for all articulators. The J displacement for /p/ was larger than /m/ for three subjects; S1, S3, S4 in the /ae/ context.

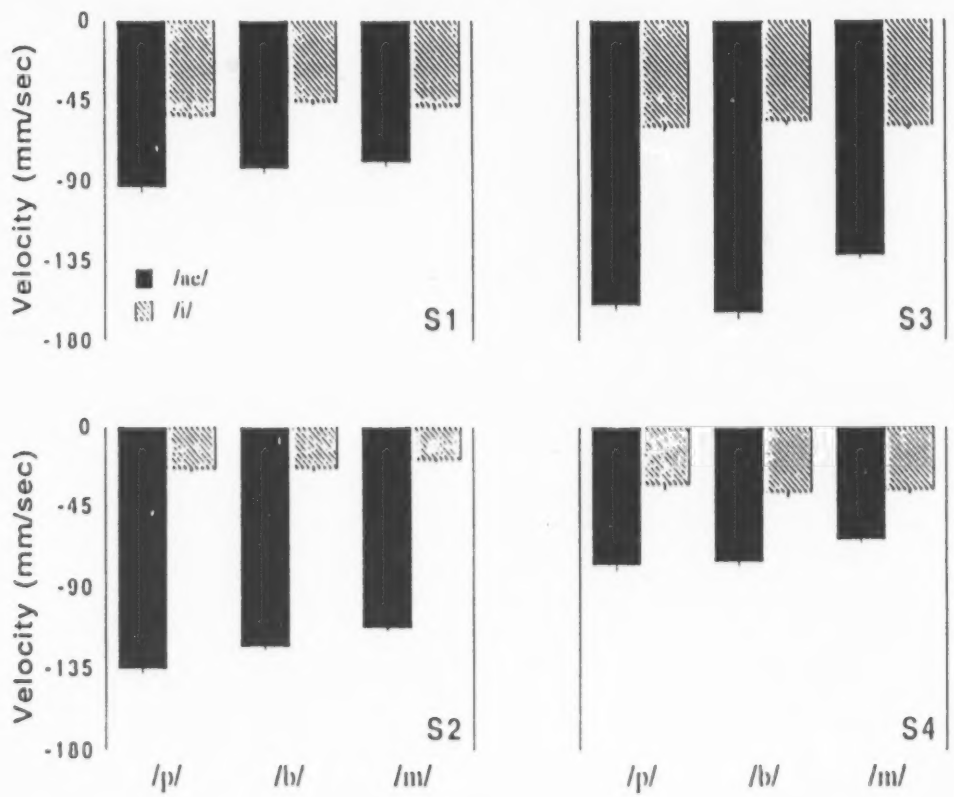


Figure 5. **Average jaw opening velocity, in millimeters per second (mm/sec), for the different phonetic contexts for each of the four subjects. Vertical bars indicate one standard error.**

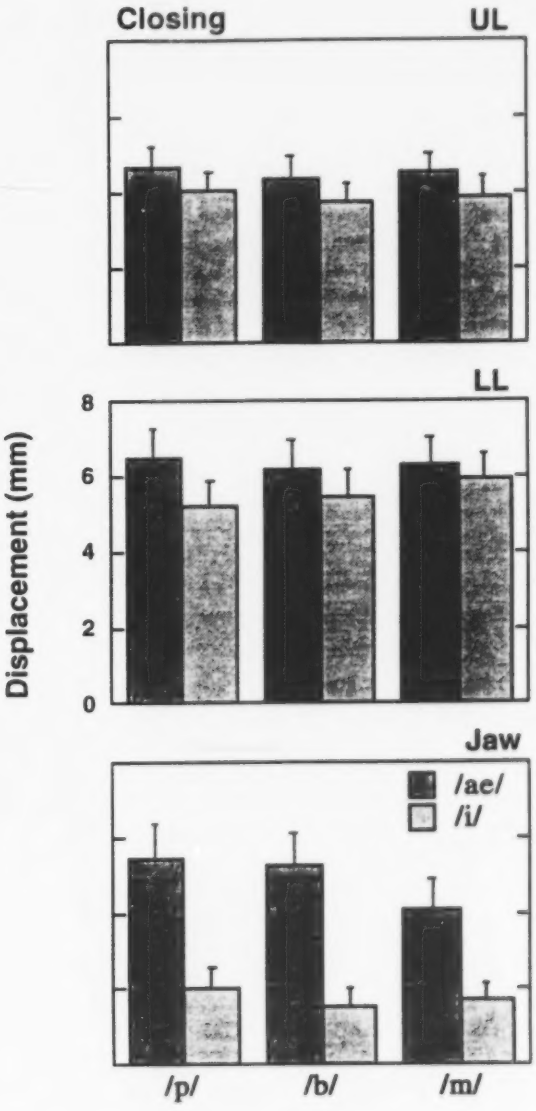


Figure 6. **Average upper lip (UL), lower lip (LL), and jaw (J) closing displacement (in millimeters) for the phonetic contexts for the group. Vertical bars indicate one standard deviation.**

Oral closing—Velocity/Stiffness

Lower lip closing velocity has previously been shown to be higher for /p/ than /b/ (Sussman et al, 1973; Summers, 1987). A similar result for the LL was found in the present investigation although the differences, with few exceptions, were small. Shown in Figure 7 is a summary of the LL closing movement velocities for all subjects for the different phonetic contexts. The closing velocity for /p/ was higher for /b/ for three subjects (S2, S3, S4) and higher than /m/ for two subjects (S3, S4) in the /ae/ context. For S1 the LL closing velocity was significantly reduced for /p/ compared to /m in the /ae/ context ($p < .01$). Table 3 is a summary of the post hoc comparisons for each articulator and subject. As shown, the UL results are generally similar to the LL with S2, S3, and S4 showing small increases in closing velocity for certain comparisons in the /ae/ context. Only two subjects (S3, S4) showed higher J closing velocities for /p/ compared to /b/ or /m/ in the /ae/ context. Jaw closing velocity for all subjects was significantly higher in the /ae/ context. A similar tendency was noted for the UL and LL although the results were not as robust as those for the J.

It appears that the magnitude of the LL closing velocity provides a fairly robust metric differentiating voiceless from voiced consonants. However, it was of interest to determine if a combination of kinematic variables, especially those for the UL and J, might provide a more reliable measure. One construct that has been used to describe speech movements and suggested as an important control variable for speech is mass-normalized stiffness, expressed as the ratio of peak velocity to peak displacement (Kelso et al., 1985; Ostry & Munhall, 1985). Changes in derived stiffness have been shown to co-occur with changes in movement duration, speaking rate and emphasis. The correlation of velocity to displacement was generally high for all subjects and articulators with the correlation magnitudes ordering J > LL > UL. Velocity/displacement correlations were generally high across subjects for the LL and J ranging from r = .49 to r = .79 for the LL and from r = .91 to r = .96 for the jaw; the upper lip was less robust with correlations ranging from r = .16 to r = .86. Presented in Figure 8 are the UL, LL, and J stiffness values for the different consonants for the group. The LL values are generally higher for /p/, a finding consistent with the peak velocity measures presented in Figure 7. For the UL and J, the ratios of peak velocity to peak displacement suggest that in the /ae/ context, /p/ is less stiff than /m/, a result opposite to that for the LL. Vowel related differences were also inconsistent across subjects with the exception that J stiffness was always higher for /i/ than /ae/. The data from the individual subjects reflected the group trend and are presented in Table 4.

Table 2. *Mean difference (x_d) and associated F-value (Scheffe's F-test) for within-vowel (/aep-aeb/, /aep-aem/, /ip-ib/, /ip-im/) consonant comparisons and across-vowel (/aep-ip/, /aeb-ib/, /aem-im/) consonant comparisons for the upper lip, lower lip, and jaw closing displacement (mm) for all subjects. Degrees of freedom for S1-3 (5, 234), S4 (5,232); asterisk indicates a significant difference at p < .01.*

Upper Lip

	/aep-aeb/	/aep-aem/	/ip-ib/	/ip-im/	/aep-ip/	/aeb-ib/	/aem-im/
S1 (x_d)	.75	.56	.2	.26	.4	-.14	.1
F	8.32*	4.61*	.61	.99	2.4	.31	.16
S2	.58	.18	.32	.4	.77	.51	.99
	6.86*	.69	2.06	3.26*	12.2*	5.33*	19.95*
S3	.-55	-.07	-.01	-.07	.51	1.04	.5
	7.26*	.11	.003	.14	6.24*	26.41*	6.09*
S4	.58	-.03	.34	.-06	1.06	.82	1.04
	1.95	.01	.67	.02	6.59*	4.0*	6.3*

Lower Lip

	/aep-aeb/	/aep-aem/	/ip-ib/	/ip-im/	/aep-ip/	/aeb-ib/	/aem-im/
S1 (x_d)	-.24	-.36	-.03	-.34	.26	.47	.27
F	.66	1.42	.01	1.33	.77	2.49	.84
S2	.06	-.28	-1.17	-1.45	2.0	.77	.82
	.02	.59	10.05*	15.49*	29.17*	4.33*	4.98*
S3	.33	.97	.08	-.21	1.76	1.51	.58
	.7	6.0*	.04	.28	19.81*	14.59*	2.17
S4	1.1	.48	.08	-.75	1.04	.03	-.18
	7.09*	1.34	.04	3.26*	6.4*	.01	.19

Jaw

	/aep-aeb/	/aep-aem/	/ip-ib/	/ip-im/	/aep-ip/	/aeb-ib/	/aem-im/
S1 (x_d)	.48	.87	.46	.2	2.47	2.44	1.79
F	1.06	3.49*	.96	.18	27.81*	27.31*	14.68*
S2	-.46	.46	.99	.28	3.93	5.38	3.75
	1.25	1.26	5.82*	.47	91.51*	171.59*	83.32*
S3	-.22	1.97	.09	.18	4.58	4.88	2.79
	.38	31*	.06	.26	167.02*	190.02*	61.92*
S4	.77	1.84	.47	.57	2.25	1.95	.98
	11.93*	68.34*	4.61*	6.61*	102.3*	78.32*	19.53*

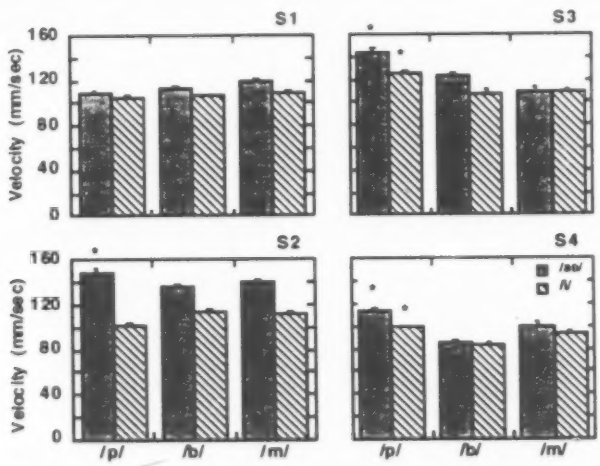


Figure 7. **Mean lower lip peak closing velocity (in millimeters per second) for the different phonetic contexts for the four subjects. Vertical bars indicate one standard error; asterisk indicate significantly higher closing velocity (p < .01) for /p/ compared to /b/ and /m/.**

Table 3. *Mean difference (x_d) and associated F-value (Scheffe's F-test) for within-vowel (/aep-aeb/, /aep-aem/, /ip-ib/, /ip-im/) consonant comparisons and across-vowel (/aep-ip/, /aeb-ib/, /aem-im/) consonant comparisons for the upper lip, lower lip, and jaw closing velocity (mm/sec) for all subjects. Degrees of freedom for S1-3 (5, 234), S4 (5,232); asterisk indicates a significant difference at p < .01.*

Upper Lip

	/aep-aeb/	/aep-aem/	/ip-ib/	/ip-im/	/aep-ip/	/aeb-ib/	/aem-im/
S1 (x_d)	4.52	-1.16	3.11	3.43	-2.18	-3.59	2.42
F	1.36	.09	.64	.78	.32	.86	.39
S2	2.26	8.6	4.34	-5.64	10.46	8.38	24.7
	.4	5.87*	1.49	2.53	8.68*	5.57*	48.43*
S3	5.96	5.92	-.27	-2.24	.24	6.47	8.4
	4.26*	4.2*	.01	.6	.01	5.02*	8.45*
S4	13.76	2.86	8.47	10.66	9.49	-4.27	17.29
	4.64*	.2	1.78	2.79	2.21	.45	7.34*

Lower Lip

	/aep-aeb/	/aep-aem/	/ip-ib/	/ip-im/	/aep-ip/	/aeb-ib/	/aem-im/
S1 (x_d)	-4.9	-10.14	-.92	-4.34	2.99	6.97	8.79
F	.79	3.38*	.03	.62	.29	1.6	2.54
S2	12.45	8.62	-13.88	-11.08	47.79	21.46	28.09
	3.43*	1.64	4.26*	2.71	50.49	10.18*	17.45*
S3	22.22	35.05	17.33	15.06	20.26	25.37	.27
	6.26*	15.57*	3.81*	2.87	5.2*	2.99	.0001
S4	23.87	13.6	15.46	4.93	14.5	1.59	5.84
	14.85*	3.41*	4.47*	.45	3.88*	.05	.63

Jaw

	/aep-aeb/	/aep-aem/	/ip-ib/	/ip-im/	/aep-ip/	/aeb-ib/	/aem-im/
S1 (x_d)	2.53	4.3	6.32	4.52	34.16	37.95	34.37
F	.11	.32	.69	.35	19.99*	24.68*	20.24*
S2	-1.33	10.76	14.1	1.82	68.36	83.81	59.44
	.03	2.26	3.87*	.06	91.11*	136.87*	68.85*
S3	-8.13	32.21	2.11	4.13	80.43	90.67	52.35
	1.43	22.53*	.1	.37	140.5*	178.56*	59.53*
S4	14.62	27.33	6.23	8.93	35.67	27.27	17.26
	18.67*	65.2*	3.43*	6.96*	111.03*	65.75*	26.02*

Table 4. *Derived stiffness values, defined as the ratio of peak velocity to peak displacement for the first oral closing movement for the upper lip (UL), lower lip (LL), and jaw (J) for the four subjects. Results are presented according to vowel and consonant context for the first movement sequence. A single asterisk indicates a significant /p/ /b/ or /m/ difference at p < .01; double asterisks indicate that /p/ was significantly different than both /b/ and /m/. All comparisons were within vowel.*

		/aep/	/aeb/	/aem/	/ip/	/ib/	/im/
	S1						
UL		13.4 (.24)**	14.8(.25)	15.5(.25)	15.1(.16)	15.1(.20)	15.2(.18)
LL		18.2(.28)	18.3(.22)	18.7(.19)	18.5(.20)	18.6(.22)	18.1(.22)
J		17.7(.20)*	19.0(.28)	20.2(.26)	20.8(.28)	22.2(.34)	20.7(.36)
	S2						
UL		12.7(.20)**	15.1(.32)	15.0(.17)	12.6(.27)*	14.7(.21)	12.4(.18)
LL		21.0(.27)*	19.4(.34)	19.0(.24)	19.8(.22)**	18.3(.21)	17.1(.25)
J		18.9(.19)	17.7(.15)	18.4(.17)	21.8(.42)**	29.4(.79)	24.2(.46)
	S3						
UL		17.0(.32)	16.6(.63)	18.7(.37)	20.4(.31)	20.3(.36)	19.0(.45)
LL		20.1(.21)**	17.9(.28)	17.6(.42)	23.2(.49)**	20.3(.42)	19.5(.42)
J		19.6(.14)	20.3(.20)	21.3(.24)	28.0(.85)	27.8(.82)	27.1(.63)
	S4						
UL		15.6(.22)	14.6(.24)	14.9(.21)	16.9(.24)*	16.3(.27)	14.7(.25)
LL		20.0(.28)	18.7(.48)	19.2(.35)	21.5(.34)*	18.4(.34)	17.5(.30)
J		16.8(.20)	16.2(.27)	18.3(.26)	17.8(.23)*	19.9(.43)	19.2(.40)

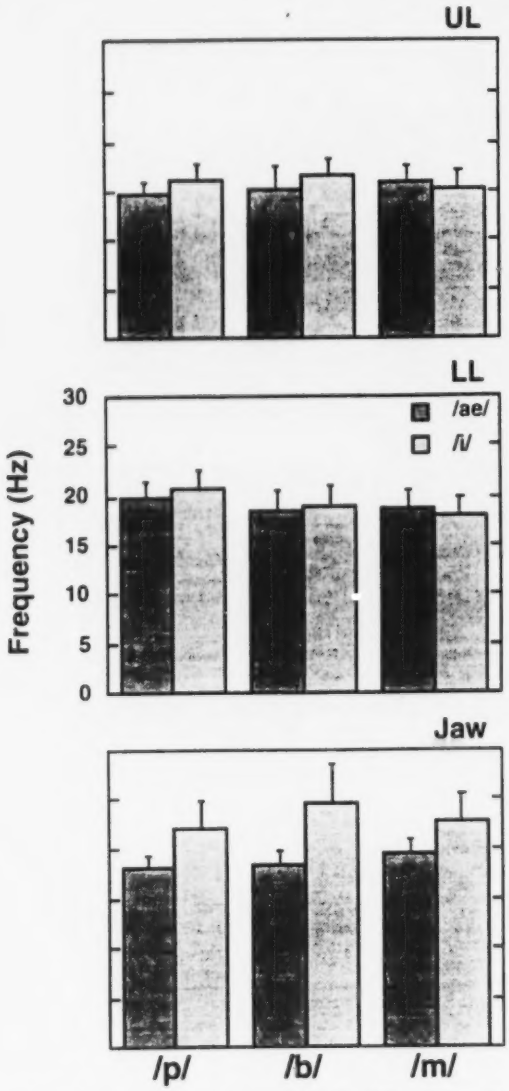


Figure 8. **Average mass-normalized stiffness for the upper lip (UL), lower lip (LL) and jaw (J) oral closing for the different vowel-consonant combinations for the group. Mass normalized stiffness was derived by dividing the peak velocity, in mm/sec, by the peak closing displacement, in mm, resulting in units of frequency (1/sec). Vertical bars indicate one standard deviation for the group.**

Opening/closing interactions

Inspection of the oral opening and closing movement relations collapsed across consonants revealed some interactions that were not phonetically related but apparently reflect more general characteristics of speech movement organization. As indicated above, the positions of the lips and jaw for the first vowel were not reliably related to phonetic context. However, vowel-related positions of the different articulators did vary and those variations influenced the extent of articulator displacement for the closing movement. Shown in Figure 9 are results from the LL (left side) and J (right side) opening position-closing displacement relations for the individual subjects. As shown, the relative articulatory positions for the vowel produced systematic variations in the magnitude of the oral closing movement. The LL data (left panel) reflect all consonant and vowel combinations; due to the large movement difference for the J, only the data for the /ae/ context are included (right panel). Including /i/ in the /ae/ data created a range effect that artificially inflated the correlation.

As shown in Figure 9, correlation coefficients for the LL range from r = .37 to r = .66. The J closing displacement was even more strongly related to the preceding J opening position with correlation coefficients ranging from r = .50 to r = .73; the UL (not shown) was less consistent (correlation coefficients ranging from r = .05 to r = .55). As indicated, the lower the position of the lips and jaw from the subjects defined rest position prior to closure, the greater the resulting displacement.

In addition to the apparent dependence of articulator displacement on articulator position, other characteristics of oral opening and oral closing were found to covary. For example, the jaw opening velocity for /ae/ and the subsequent closing velocity for the consonant demonstrated a covariation with correlations ranging from r = .47 to r = .67 for the four subjects (Figure 10). It should also be noted that there was a trend for jaw opening and closing velocity for /p/ and /b/ to be faster than for /m/. It appears that there are systematic spatiotemporal variations in oral opening and closing and such interactions are not necessarily phoneme specific.

Articulator interactions. In previous studies it has been suggested that the movement of individual articulators are subordinate to the combined movement of the contributing parts (Hughes & Abbs, 1976; Gracco & Abbs, 1986; Saltzman, 1986). That is, there appears to be a higher level motor plan in which the action of individual articulators is partially dependent on the action of the other articulators contributing to a multiarticulator goal. In the present study, separate stepwise regressions were done on the individual UL, LL, and J displacements for oral closing, using the positions of the three articulators for the preceding vowel as independent variables, to examine for evidence of articulatory interactions.

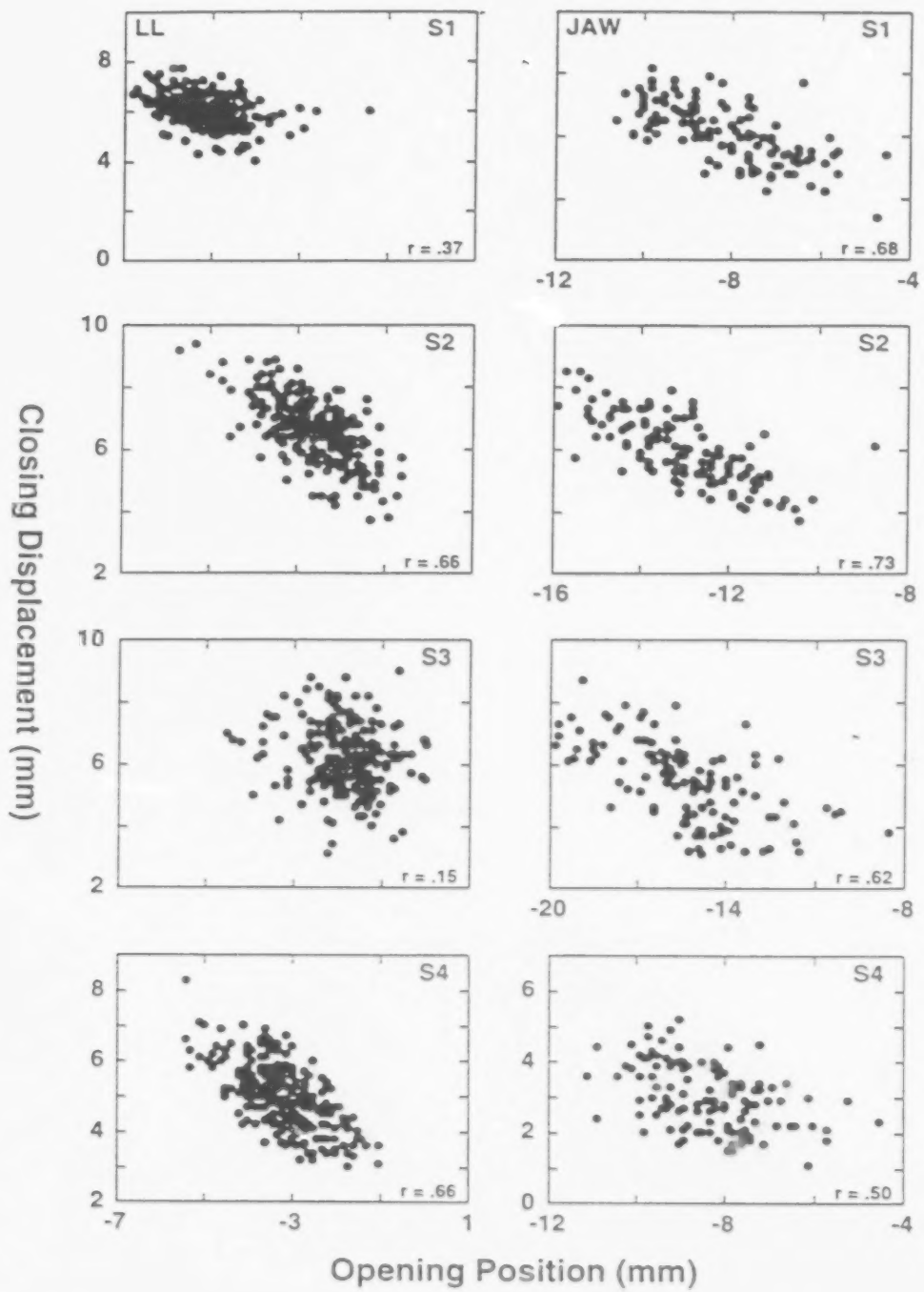


Figure 9 . Scatter plots of the lower lip (LL; left side) and jaw (Jaw; right side) closing displacements as a function of the relative positions of the respective articulators for the preceding vowel . Product-moment correlation coefficients (r) are presented at the bottom right hand corner of each plot. Only the jaw opening positions/jaw closing displacements are presented for /ae/ (see text for explanation).

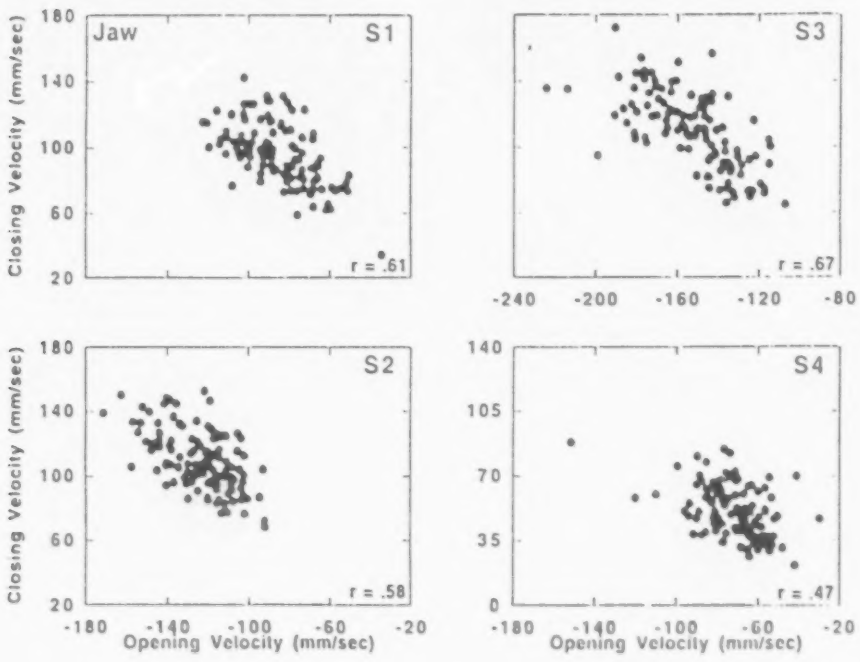


Figure 10. **Jaw opening peak velocity for /ae/ and the corresponding jaw closing peak velocity for the three consonants. Product-moment correlation coefficients (r) are presented at the bottom right hand corner of each plot. As shown, as the jaw opens faster there is a tendency for the jaw to close faster as well.**

In all cases, the position of a particular articulator for the preceding vowel was found to have the strongest influence on the movement displacement for this articulator. However, in all cases significant increases in the amount of explained variance were noted when the position of at least one other articulator was included in the regression model. A summary of the regression results is presented in Table 5.

Second sequence

Cycle duration. In the present investigation, the stimuli required two complete movement sequences; two opening and two closing. Following the first sequence in which the vowel and consonant varied, the remaining sequence for "apple" was similar in phonetic context; an opening for the vowel /ae/ and a closing for /p/. Thus the data could be examined for carryover effects of the first syllabic context on the second movement sequence. Examination of the duration of the second movement cycle presented in Figure 11 revealed significant differences as a function of the preceding consonant. For subjects S1, S3 and S4, the duration of the second movement sequence, obtained from the lip aperture signal, was longer when the preceding oral closing consonant was /p/ than when it was /b/ or /m/ (S1 [$F = 52.06$, $p = .0001$]; S3 [$F = 42.23$, $p = .0001$]; S4 [$F = 49.13$, $p = .0001$]). There were no consistent vowel related effects for these subjects ($p > .05$). For S2 the duration of the second movement sequence was only longer for /p/ when the preceding vowel was /i/ ([$F(5,234) = 6.92$, $p < .01$]); a significant vowel effect was noted in the /p/ context ($F = 14.06$, $p < .01$).

To determine whether the longer cycle duration was localized to a single phase of the movement sequence and a single articulator, or distributed across the opening and closing phases and contributing articulators, the movement characteristics of the respective phases for all three articulators were examined. Shown in Figure 12 are the group averaged upper lip, lower lip, and jaw opening and closing movement durations for the second movement sequence. As can be seen, the opening movements account for the changes in cycle duration associated with the preceding phonetic context noted in Figure 11. For the group the opening movement duration was longer following /p/ than /b/ or /m/ for the UL ([$F(2,956) = 47.07$, $p = .0001$]) the LL ([$F(2,956) = 95.9$, $p = .0001$]) and the J ([$F(2,956) = 31.53$, $p = .0001$]). There were no

significant closing movement duration changes for any articulator or context ($p > .05$). In addition, vowel related differences were noted that were obscured by examining the sequence duration from the lip aperture signal. For the UL and LL the opening movement was shorter for /i/ compared to /ae/ ([$F(1,957) = 212.02, 125.01$, $p = .0001$] for the UL and LL respectively). For the J just the opposite was found; jaw opening duration was longer for /i/ compared to /ae/ ([$F(1,957) = 160.45$, $p = .0001$]). Opposing vowel-related effects for the lips and jaw apparently offset one another in the combined lip aperture signal resulting in no net change to the second movement cycle.

Table 5. *Stepwise regression of the upper lip (ul), lower lip (ll), and jaw (j) oral closing displacement (disp) for the first opening/closing sequence. The displacement of each individual articulator was regressed on the relative position of each of the three articulators for the preceding vowel.*

	Upper Lip	**R**	**adj. R-squared**
S1	UL disp = 3.45 + .56(ul) +.15(j) +.27(ll)	.66	.42
S2	UL disp = 3.06 + .38(ul) + .08(j)	.37	.12
S3	UL disp = 1.45 + .64(ul) + .13(j) + .09(ll)	.67	.43
S4	UL disp = 2.02 + .84(ul) + .21(j) + .2(ll)	.67	.44
	Lower Lip	**R**	**adj. R-squared**
S1	LL disp = 2.63 + .37(ll) + .26(j) + .26(ul)	.67	.44
S2	LL disp = 2.66 + .71(ll) + .18(j)	.72	.50
S3	LL disp = 3.99 + .32(ll) + .14(j)	.36	.11
S4	LI disp = .1 + .6(ll) + .28(j) + .46(ul)	.78	.60
	Jaw	**R**	**adj. R-squared**
S1	J disp = .85 + .58(j) + .66(ul)	.74	.53
S2	J disp = 3.1 + .61(j) + .32(ul) + .22(ll)	.76	.57
S3	J disp = .07 + .37(j) - .49(ul)	.64	.40
S4	J disp = .01 + .48(j)	.50	.24

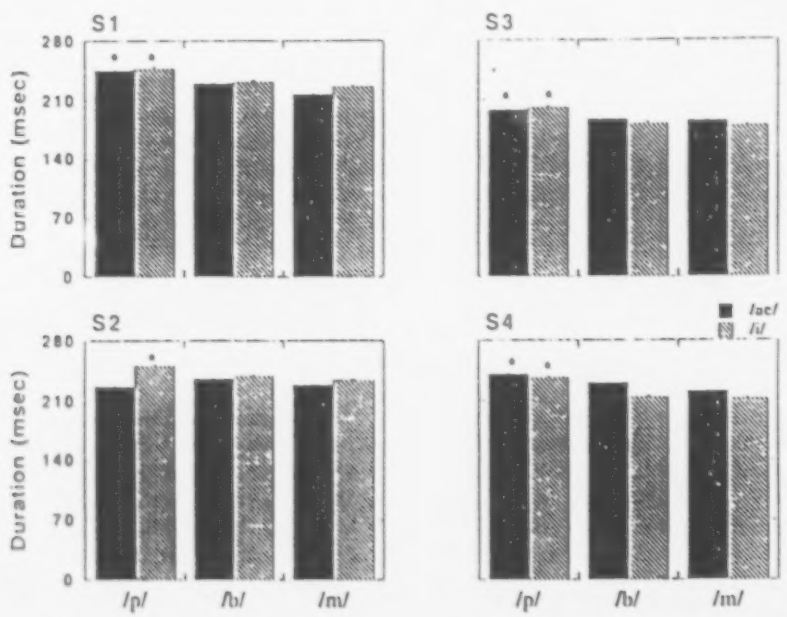


Figure 11. Duration (in milliseconds) of the second movement cycle (S-2) for the oral opening for /ae/ and subsequent closing for /p/ following the different vowel-consonant combinations in the first cycle (S-1) for each of the four subjects. Vertical bars indicate one standard error; asterisks indicate significant /p/ - /b/ or /m/ difference ($p < .01$).

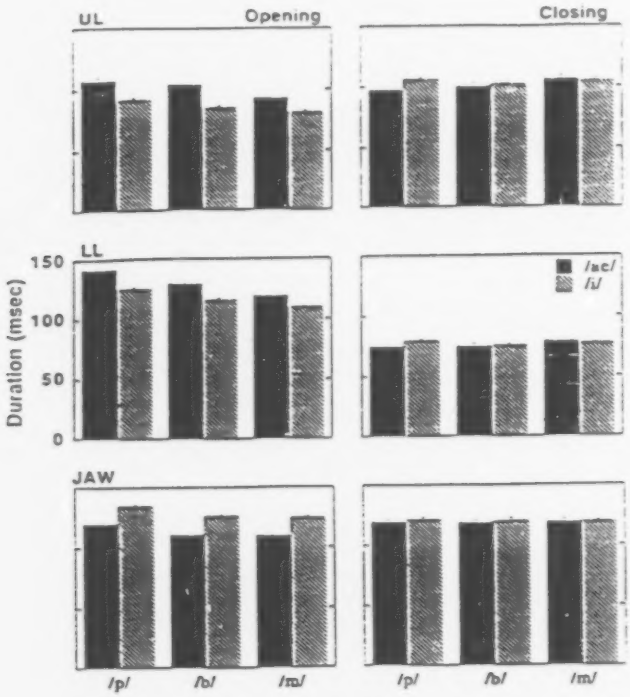


Figure 12. Group means of the duration of opening (Op_2) and closing (Cl_2) movements for the second sequence for the upper lip (UL), lower lip (LL) and jaw. The opening movement durations for the UL and LL were reliably longer for /p/ compared to /b/ and /m/. Vertical bars indicate one standard error.

Velocity/Displacement

Other characteristics of the opening phase of the second sequence revealed articulator-specific contextual differences in movement extent and movement velocity consistent with these results. In the opening phase, the most significant effects were vowel related for the J and consonant related for the LL. As summarized in Figure 13 for the four subjects, the J opening movement for the /ae/ went farther (S1-S3[F(1,238) = 179.72, 65.1, 709.07 respectively, p = .0001]; S4[F(1,237) = 298.2, p = .0001]) and faster when following /i/ (S1-S3[F(1,238) = 78.06, 12.01, 419.24 respectively, p = .0001]; S4[F(1,237) = 306.88, p = .0001]). The effect for the LL was less robust (displacement; S3 F = 290.46, p = .0001; S4 F = 52.05, p = .0001; velocity S2 F = 6.93, p = .009; S3 F = 21.7, p = .0001) and when significant went in the opposite direction to that observed for the J. The greater J opening displacement, duration and velocity for /ae/ following /i/ appears to be due to the higher jaw position prior to the second opening movement following /i/. The jaw position for /i/ is higher than when the first vowel is /ae/ and as a consequence, the jaw opening displacement starts from a significantly higher position. The higher J position results in larger opening movement displacements, longer opening movements and higher opening velocities for the subsequent vowel opening.

In contrast to the vowel related opening differences, consonant related differences were noted but only for the LL. The magnitude of the opening velocity following closure was found to order according to consonant identity with /p/ < /b/ < /m/ for all subjects. As can be seen in Figure 14, this result was reliable for all subjects for the LL (S1-S4 F = 52.95, 237.76, 54.44, 160.85 respectively, p = .0001). For the J post hoc testing revealed only two significant differences; /b/ - /m/ for S1 (F = 6.01, $p < .01$) and /p/ - /m/ for S2 (F = 31.52, $p < .01$).

Lip-Jaw Coordination

Oral closing. Starting from a relatively steady state for the /s/ in the six stimulus words, the upper lip, lower lip, and jaw attained some open posture for the vowel and subsequently closed the oral opening cooperatively. Using the jaw opening peak velocity for the vowel as a reference point

(see Figure 2), the consistency of the relative timing of the three articulators was examined. Consistent with previous studies the timing of the UL, LL, and J covaries during oral closing (Gracco, 1988; Gracco & Abbs, 1986). The UL-LL relative timing for the four subjects is presented in Figure 15 with the corresponding correlations shown at the bottom. Similar results were obtained for the LL-J with all within-vowel correlations ranging from r = .86 to r = .99.

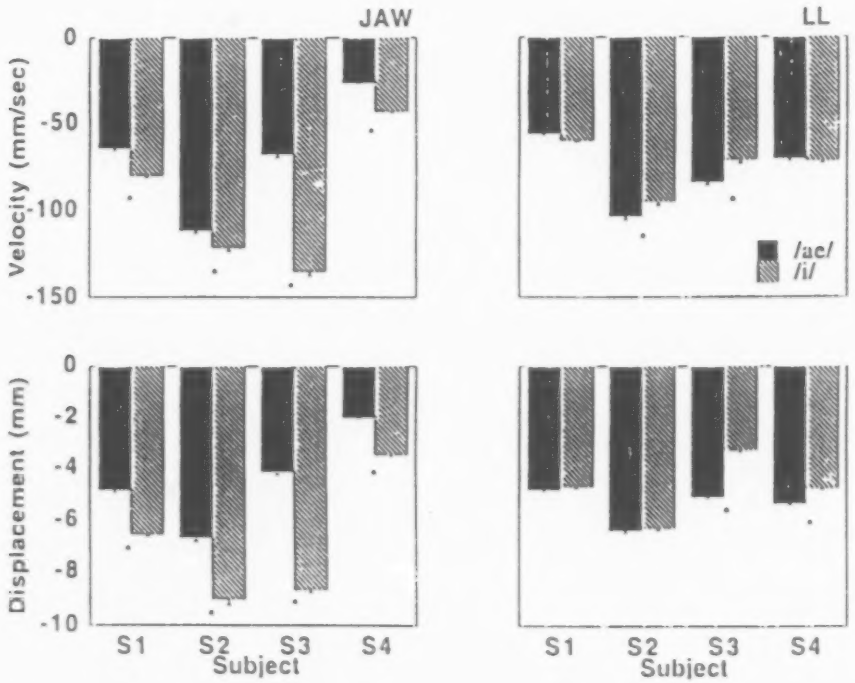


Figure 13. **Peak opening velocity and displacement for jaw (Jaw) and lower lip (LL) for the second movement sequence for the four subjects. Vertical bars indicate one standard error; asterisks indicate significant vowel differences ($p < .01$). As shown, the jaw opening is faster and farther when the preceding vowel is /i/ compared to /ae/. The same trend is not seen in the lower lip.**

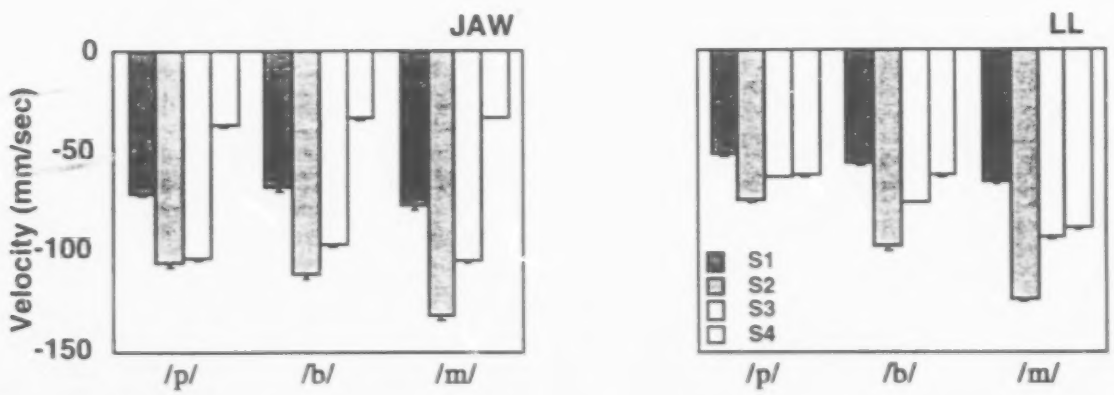


Figure 14. **Jaw and lower lip opening velocity for the second movement sequence as a function of preceding consonant identity for the four subjects. For all subjects, the LL opening velocity was highest for /m/ compared to /p/ or /b/; the same trend was not found for the Jaw. Vertical bars indicate one standard error.**

Oral opening. In contrast to the UL, LL, and J timing relations for the oral closing, the timing of the three articulators for oral opening was less consistent. Figure 16 presents the UL-LL relative timing for oral opening for the four subjects. For these data the UL-LL opening velocity timing is referenced to the time of the preceding UL peak closing velocity. While it appears that the two lips are certainly related in their timing, they do not display the same consistency seen for the oral closing (Figure 15). Similar results were obtained for the LL-J with correlations ranging from r = .57 to r = .80.

Context effects. A final issue related to the lip-jaw coordination for oral closing focused on whether the coordinative relations among the articulators are fundamentally the same or different for the different contexts. To examine the articulatory relations in detail, the intervals between the UL-LL and LL-J peak velocities were examined.

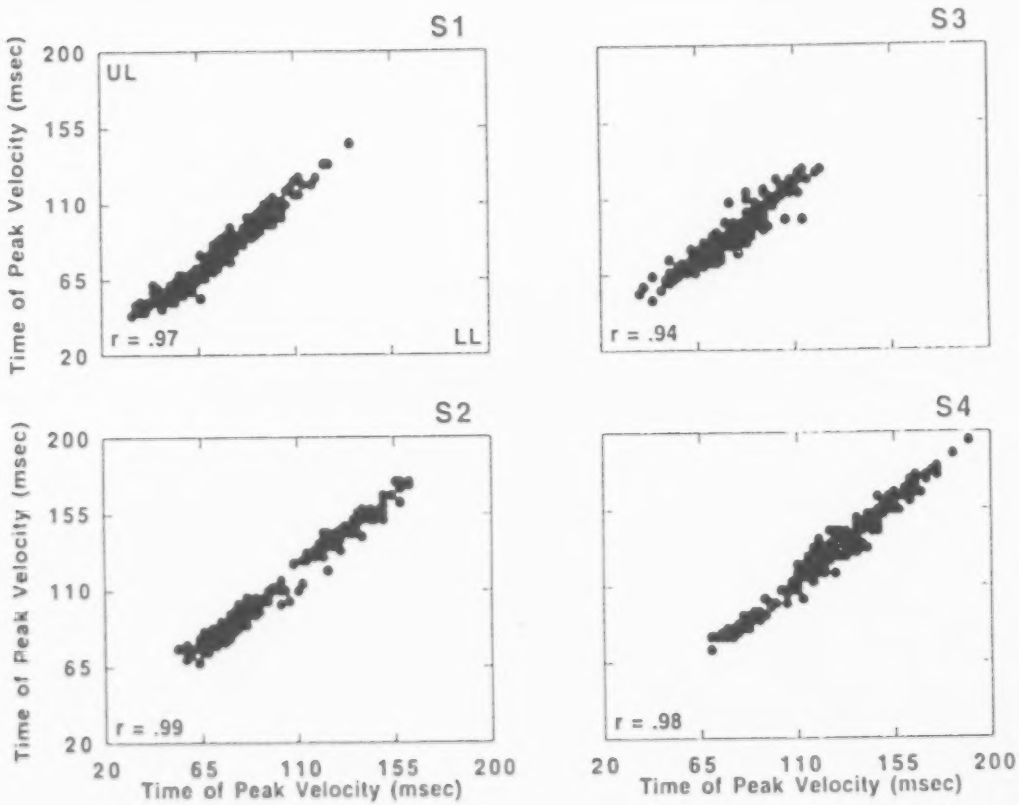


Figure 15. Scatter plots of the relative timing of the upper lip (UL) and lower lip (LL) peak closing velocity (Cl_1) for the different vowel-consonant contexts for the four subjects. Time of peak velocity was referenced to the time of the jaw opening peak velocity for the first vowel (Op_1). Product-moment correlation coefficients (r) are presented at the bottom right hand corner of each plot.

The interpeak interval is a measure of the absolute time difference between articulator pairs and the sign of the difference (positive or negative) indicates whether one articulator leads or lags another. Figure 17 presents the averages of the UL-LL (right side) and LL-J (left side) interpeak intervals for oral closing for the four subjects. The most significant effects were vowel-related. For the UL-LL, the tendency was for the interpeak interval to decrease in the /i/ context with four of the twelve possible within-vowel comparisons (3 comparisons X 4 subjects) reaching significance (S2 F = 6.88 for /b/; S3 F = 4.25 for /p/; S4 F = 5.32 and F = 9.91 for /p/ and /b/ respectively, $p < .01$). The LL-J interval also displayed vowel related effects for all subjects, however, the trend was opposite of that seen for the UL-LL. Of the twelve possible differences, five were found to be reliable (S2 F = 3.82 and 3.82 for /b/ and /m/ respectively; S3 F = 6.12 for /m/; S4 F = 20.66 and 4.52 for /p/ and /m/ respectively, $p < .01$). For the /i/ context, the LL-J interval tended to increase indicating that the J timing was not being adjusted to the vowel context. Consonant related effects for the UL-LL were essentially absent with only two /p/ comparisons reaching significance (S3 F = 8.51 and S4 F = 5.5, $p < .01$). For the LL-J, there was a trend for the interval for /p/ closure to be longer than /b/ and/or /m/ with reliable differences found for three subjects (S2 F= 18.99 for /p/ - /b/ in the /ae/ context and F = 45.27 in the /i/ context; S3 F = 7.48 for /p/ - /m/; S4 F = 14.36 for /p/ - /b/ in the /ae/ context and F = 48.29 for /p/ - /b/ in the /i/ context; $p < .01$). Again it appears that the J timing is less related to the phonetic context than are the lips.

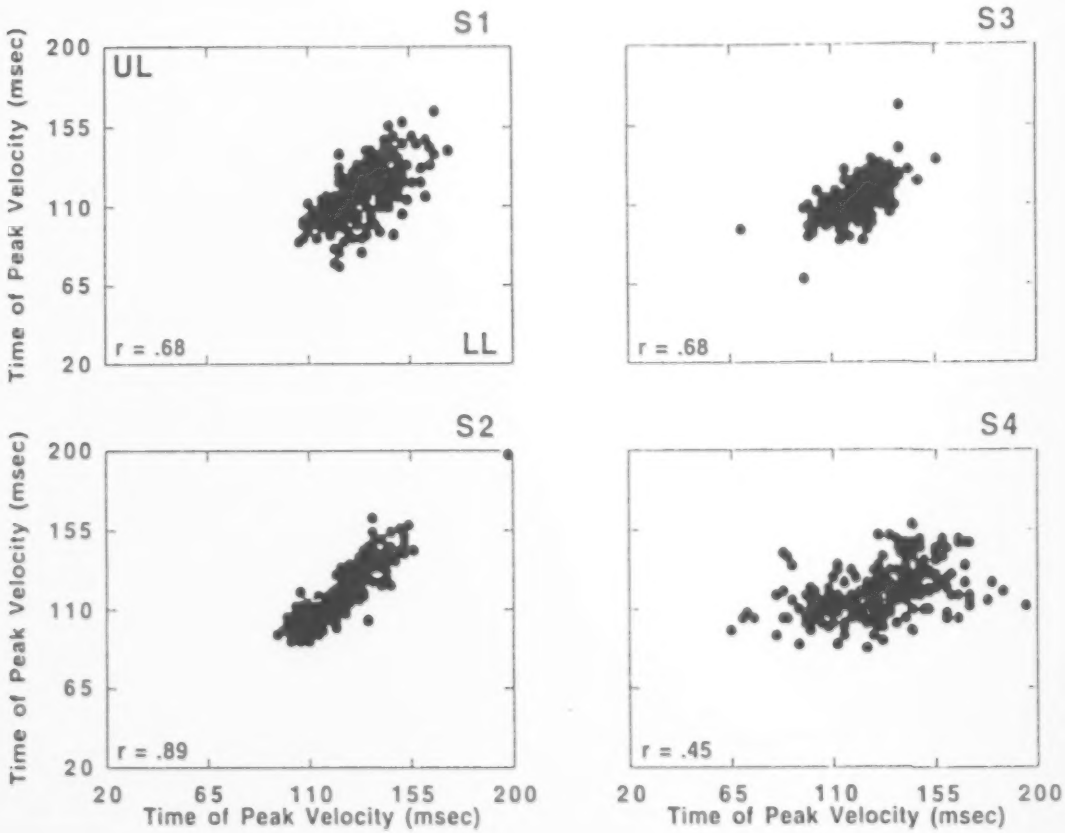


Figure 16. Scatter plots of the relative timing of the upper lip (UL) and lower lip (LL) peak velocity for the second oral opening (Op_2) for the four subjects. Time of peak velocity was referenced to the time of the upper lip peak velocity for the oral closing (Cl_1). Product-moment correlation coefficients (r) are presented at the bottom right hand corner of each plot.

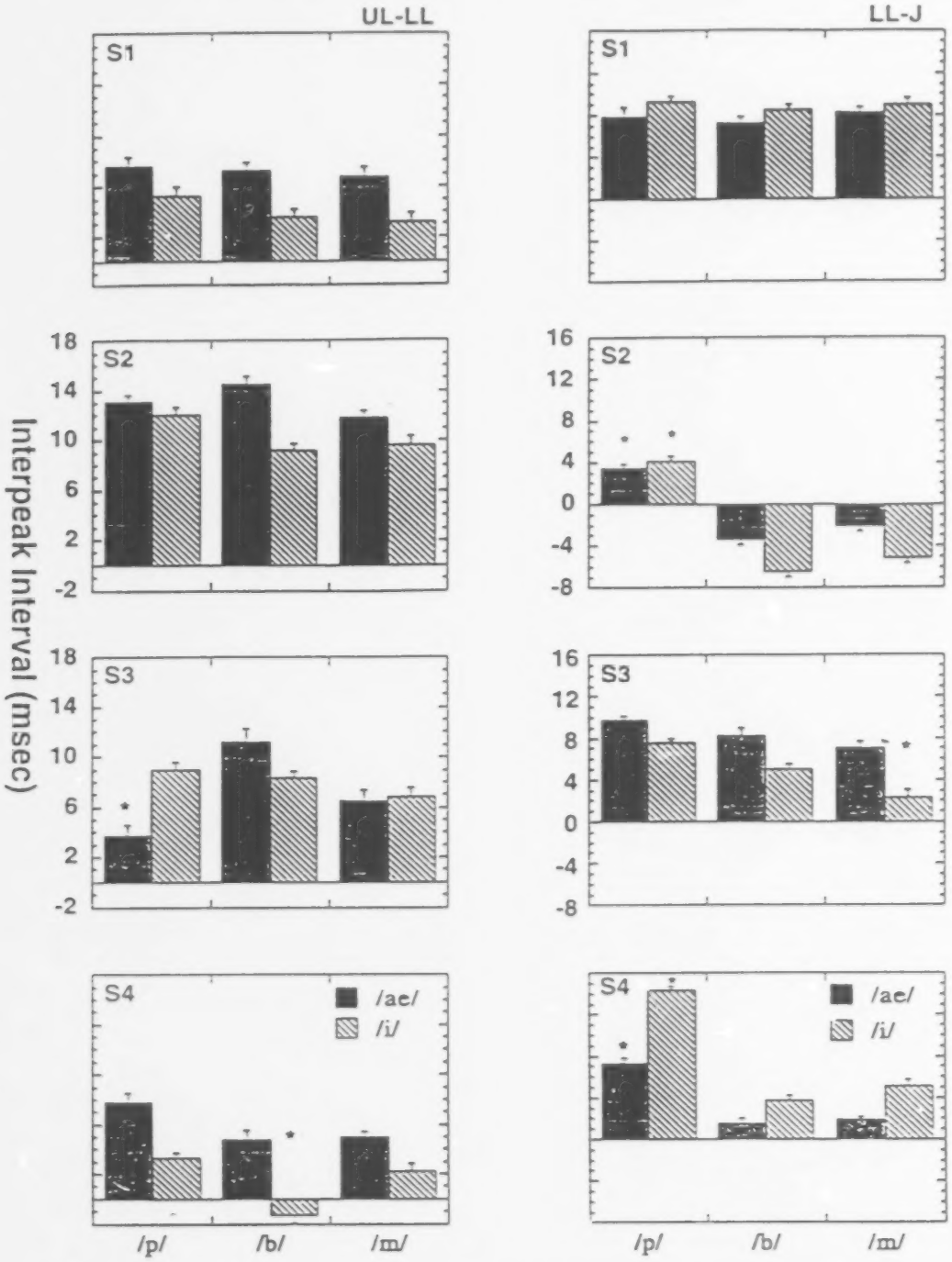


Figure 17 . **Upper lip-lower lip (UL-LL; left side) and lower lip-jaw (LL-J; right side) closing (Cl_1) velocity interpeak intervals for the four subjects The interpeak interval was defined as the time, in milliseconds, between the occurrence of the peak velocities for the different articulator pairs. A single asterisk centered over either /p/ bar indicates a single significant within-vowel consonant comparison; a double asterisk indicates that both comparisons (/p/ - /b/ and /p/ - /m/) reached significance ($p < .01$). An asterisk centered over any consonant pair indicates a significant within-consonant vowel comparison ($p < .01$).**

It can also be seen that the sequence of velocity events varied across subjects and articulator pairs. Positive values for the two interpeak intervals indicate that the average sequence of articulator velocity peaks was UL-LL-J. For S2 and S4, negative values were obtained for /b/ and /m/ for the LL-J and UL-LL respectively indicating that the sequence of events reversed from UL-LL-J to either UL-J-LL or LL-UL-J. In general, the UL-LL timing for the different phonetic contexts suggest that the coordinative relations among these articulators remains similar for different consonants. When the movement characteristics are substantially modified, as occurred for the different vowels, there is a tendency for the relative timing of the articulators to change with differences noted for the lips.

DISCUSSION

The present study investigated characteristics of the lip and jaw movements associated with a class of consonants (bilabials). A number of observations were made that reflect on certain speech movement principles and the manner in which such principles are modulated for phonetic context. The most consistent finding of the present investigation is that the timing or phasing of contiguous movement phases is an important mechanism for modifying speech movements for certain phonetic differences (see also Browman & Goldstein, 1989; 1990; Edwards, Beckman, & Fletcher, 1991; Saltzman & Munhall, 1989). Speech movements appear to be organized and hence controlled at a level in which multiple articulatory actions and contiguous movement phases are the functional units of control and coordination. These conclusions will be discussed below.

Movement adjustments. The two vowels used in the present study resulted in articulatory changes in position, extent and speed due to greater movement requirements for the jaw (Lindblom, 1967; Lindblom, Lubker, & Gay, 1979). As such, the size of the oral opening, determined predominantly by jaw position, is a significant factor in determining movement adjustments for different vowel contexts. In contrast, the upper and lower lips were not affected by the identity of the two vowels. Similar findings have been reported by Macchi (1988) in which lip position for /p/ was not affected by the vowel environment while the jaw was due to tongue height. It should be noted that transducing lip motion in only one dimension (inferior-superior) limited any observations of anterior-posterior lip spreading, often assumed to accompany /i/. However, while there may have been some differences in the anterior-posterior dimension, it seems clear that lip motion is not the major factor in differentiating the vowels /ae/ and /i/.

The vowel-related jaw opening movements resulted in systematic and distributed differences in oral closing movement characteristics. As jaw opening was modified for the high or low vowel, the extent and speed of lip and jaw oral closing characteristics were similarly adjusted; vowel-related oral closing movement adjustments were distributed to all the contributing articulators. These vowel-related adjustments suggest that oral opening and oral closing are not independent events. Rather, oral closing actions are dependent on characteristics of the preceding oral opening action. Other evidence for systematic movement relations between oral opening and oral closing can be found in results reported by Folkins and Canty (1986), Folkins and Linville (1983), Hughes and Abbs (1976) and Kozhevnikov and Chistovich (1965).

Some of the opening/closing interactions observed were apparently consonant-related such as higher opening and closing velocities for the voiceless consonant /p/. Jaw opening velocity for /ae/ was faster when the subsequent closing movement was the voiceless /p/ compared to /b/ or /m/ (Summers, 1987). In addition, LL closing velocity was generally higher for /p/ than /b/ or /m/ independent of the preceding vowel, and, LL opening velocity following closure, was always lowest for /p/ and highest for /m/. Previous investigations have demonstrated faster lower lip or jaw closing velocity for voiceless compared to voiced consonants (Chen, 1970; Fujimura & Miller, 1979; Sussman et al., 1973; Summers, 1987), and slower lip opening velocity for the voiceless consonant (Sussman et al., 1973). It has been suggested that the higher LL closing velocity is to accommodate the higher intraoral pressures for /p/ compared to /b/ (Chen, 1970; Sussman et al., 1973). That is, higher intraoral air pressure may require greater lip force, manifest as higher lip closing velocity, to maintain lip contact during the closure interval. While it is generally observed that oral pressure for a voiceless sound is higher than for its voiced counterpart which is higher than its voiced nasal counterpart (Arkebauer, Hixon, & Hardy, 1967; Black, 1950), empirical evidence suggests that higher lip closing velocity does not directly translate into greater lip contact force. Using a miniature pressure sensor, Lubker and Paris (1970) were unable to find reliable differences in lip con-

tact pressure for /p/ and /b/. An additional consideration is the higher lower lip opening velocity for /m/. If oral pressure and lip velocity covary, then the lowest oral pressure sounds, like /m/, should consistently have the lower closing and opening velocity. However, in the same vowel context, lower lip opening velocity following /m/ closure was higher than /p/. It seems premature at best to conclude that lip closing velocity is modulated according to oral pressure characteristics especially when a direct measure of lip contact does not support such an interpretation.

Similarly, mass-normalized stiffness, defined as the ratio of the peak closing velocity and peak displacement, was not found to consistently differentiate the different vowel and consonant combinations. While the LL stiffness estimates for the closing movements were found to be consistently higher for /p/ than /b/ or /m/, the upper lip and jaw estimates were not. For the upper lip, estimates of mass-normalized stiffness often varied reciprocally with the LL values while the jaw values showed no consistent patterns. It is hard to reconcile a differential modulation of stiffness for three articulators contributing to the same movement or phonetic goal. While stiffness may be a convenient means of describing the motion or rate of an articulator, it is not clear that it qualifies as a controlled variable. Rather than a single physical control variable it appears that speech movements are more likely organized according to principles that take into account the function of the action related to the overall goal of communication (Gracco, 1990; 1991).

A potential explanation for voice/voiceless differences can be presented from examining the relative timing of the opening and closing actions. The relative timing or phasing of articulatory motion was found to be a consistent variable differentiating the voiced and voiceless consonants. For the voiceless /p/, the closing action was initiated earlier than for /b/ or /m/ while for oral opening, /p/ was initiated later than either /b/ or /m/. These two timing results are consistent with previous acoustic results related to durational differences for voiced and voiceless consonants. In many languages, two related durational phenomena have been reported for voiced and voiceless sounds. Vowel length is shorter and duration of the consonant closure, defined acoustically, is longer for voiceless than for voiced consonants. In the present phonetic context, the interval between jaw opening velocity for /ae/ and oral closing velocity is essentially identical to the duration of the vowel (McClean, Kroll, & Loftus, 1990). As such, the earlier onset of the closing action for the /p/ is a kinematic manifestation of the differential vowel length effect (Denes, 1955; House & Fairbanks, 1953). Similarly, the longer opening movement duration following /p/ closure is consistent with longer acoustic closure duration reported for voiceless sounds. It seems, from the acoustic durational effects reported previously and the kinematic effects reported here and elsewhere, that voiced and voiceless consonants differ primarily in their timing with other kinematic variations, such as velocity differences, secondary. The movement velocity changes may be considered a natural consequence of the timing requirements for the different consonants.

However, an explanation of articulatory timing as a means to differentiate sounds within a class leaves open the question as to why articulatory timing differs. Recently it has been suggested that /p/ - /b/ differences in vowel duration reflect a perceptually salient feature of speech production which enhances the acoustic differences between voiced and voiceless consonant sounds (Kluender, Diehl, & Wright, 1988). That is, voiced and voiceless consonants differ in their timing because of the perceptual enhancement the durational differences provide to the listener. A recent investigation by Fowler (1992), however, suggests that there is no auditory-perceptual enhancement to be obtained from the shorter vowel durations and longer consonant closure durations for /p/ versus /b/. A simpler, and less speculative explanation is that timing differences among the bilabial consonants reflect modifications to accommodate concomitant articulatory actions. In the case of /p/, the longer closure duration compared to /b/, is to accommodate the laryngeal adjustment associated with devoicing. As such, from a functional perspective /p/ is an inherently longer sound than /b/. The shortening of the vowel reflects an intrusion of the phonetic gesture for /p/, including the laryngeal devoicing, on the vowel. The closing must be faster to integrate the lip and the larynx actions while the opening is slower to provide the appropriate voice onset time. It is suggested that all voiced and voiceless sound pairs (cognates), are fundamentally differentiated by their relative timing or phasing which is a direct consequence of the supralaryngeal-laryngeal interaction and consistent with aerodynamic requirements (Harris et al., 1965; Lisker & Abramson, 1964; Lubker & Parris, 1970).

Speech movement coordination. Another robust finding of the present investigation was the

consistency of the upper lip-lower lip timing relations associated with the different bilabial consonants. For all subjects, the temporal relations of the velocity peaks for the upper and lower lips were essentially constant across phonetic contexts, indicating a constraint on their coordination. While the UL-LL timing was relatively consistent, the sequence of events (UL-LL-J) did not remain constant as had been observed in more limited phonetic contexts (Caruso et al., 1988; Gracco, 1988; Gracco & Abbs, 1986; McClean et al., 1990). However, based on a number of investigations that have reported different sequences, it is not clear that the actual pattern or sequence of articulatory events is of importance (DeNil & Abbs, 1991). Rather, it is the general trend for covariation that is indicative of an underlying invariant control parameter (Gracco, 1990; 1991). That is, while the sequence of events may change with rate, stress or phonetic context, such variation does not indicate a lack of invariance as suggested recently by DeNil and Abbs (1991). As pointed out by von Neumann (1958) the nervous system is not a digital device in which extremely precise manipulations are possible but an analog device in which precision is subordinate to reliability. Establishing the presence of invariance on fixed stereotypic patterns, similar to those observed in a mechanical system, is inconsistent with principles of biological systems. Rather, constraints on coordination are most likely specified stochastically with specific limits on variation contingent on the overall goal of the action (see also Lindblom, 1987 for arguments for the adaptive nature of variability). As suggested by the quantal nature of speech (Stevens, 1972; 1989), there are vocal tract actions that may require relatively more precise control than others because of the potential acoustic consequences. However, it is doubtful that the timing of bilabial closure is a context in which absolute precision is required (see also Weismer, 1991).

More importantly, in the present study as in previous investigations of speech movement timing, the consistent relative timing relations among the UL-LL-J indicate that the timing of these articulators is not controlled independently. Rather, timing adjustments covary across functionally related actions and contributing articulators are effectively constrained to minimize the degrees of freedom to control (Gracco, 1988). While performance variables such as phonetic context, speaking rate and the production of emphasis will change the surface kinematics, the underlying principles governing their coordination remains unchanged. Moreover, the articulator interactions such as the lip adjustments to changes in jaw position and the opening and closing velocity covariation are additional manifestations of the interdependency of speech movement coordination.

Opening/Closing differences

The present results also suggest that opening and closing movements are fundamentally different actions operating under different constraints (Gracco, 1988). Oral opening was found to be the locus of the most significant consonant and vowel related articulatory adjustments. Oral closing adjustments, on the other hand, varied primarily as a function of oral opening. Only the LL peak closing velocity for the different consonants demonstrated any consistent change across subjects. In contrast to the tightly coupled lip and jaw timing during closing, oral opening coordination was found to be more variable. The different spatiotemporal patterns observed suggest that opening and closing phases of sequential speech motor actions are differentially modifiable for context. Oral closing is generally faster, involving relatively abrupt constrictions or occlusions in various portions of the vocal tract. Oral opening is generally slower, involving resonance producing events associated with vowel sounds. Hence these two classes of movements reflect fundamentally different speech actions with distinct functional (aerodynamic and acoustic) consequences (see also Fowler, 1983; Perkell, 1969).

Why this might be so may, in part, reflect the different biomechanical influences on the closing and opening actions. Oral opening involves temporarily moving the lips and jaw from some rest or neutral position to a position that requires stretching the associated tissues (skin, muscle, ligament). For closing, the lip and jaw motion is assisted by the elastic recoil from the opening stretch. A consequence of the different biomechanical influences on opening and closing would be that opening movements could be controlled directly by agonistic muscle actions. That is, changes in lip and jaw opening may result from direct modification of jaw and lip opening muscle activity. In contrast, to counteract the elastic recoil of the lip and jaw tissue, oral closing adjustments would require some combination of reduction in agonistic activation and/or agonist-antagonist co-contraction. It is suggested that the opening action would be an easier task to control than the closing action and that the biomechanical in-

fluences would differentially affect the two actions. The closing action would be more rapid due to the contribution of tissue elasticity and less variable. Such considerations are possible factors influencing the patterns observed in the present investigation reflecting biomechanical optimizations and may account for certain aspects of the ontogeny of the phonological system.

Articulator differences—Consonants and vowels

In addition to the different functions of the opening and closing actions, it can also be suggested that the lips and jaw contribute to sound production in different ways. Examination of the lip and jaw movement characteristics suggest that while all three articulators contribute to the general opening and closing, their roles are not identical. The lower lip closing and opening velocity varied as a function of the different consonant sounds. The upper lip, in contrast did not. In addition, the lower lip and jaw closing displacements were significantly and systematically related to the opening position for the preceding vowel; the upper lip closing movement displacement was only weakly related. It is plausible that the upper lip contributes in a more stereotypic manner to consonant actions while the lower lip and jaw provide more of the details related to phonetic manipulations. Each lip contributes in an interdependent but not redundant manner to the achievement of oral closing.

From the lip and jaw differences noted in the present investigation it can be suggested that consonant and vowel adjustments are articulator specific. The jaw, while assisting in the oral closing, was more significantly involved in the production of the vowels than was either of the lips. This was noted not only for the first syllable in which the vowel varied between a phonetically high /i/ and low /ae/, but in the second syllable in which jaw motion for the same syllable varied dependent on the identity of the preceding vowel. The same pattern was not observed for the LL motion. In contrast, consonant-dependent opening movement differences were noted for the LL but were not found for the J. These results suggest that boundaries between phonetic segments and/or phonemic classes such as consonants and vowels, may be identified by differential articulatory actions that overlap in time. The speech mechanism can be thought of as a special purpose device in which individual components contribute in unique and complementary ways to the overall aerodynamic/acoustic events.

Speech motor programming/Serial timing

Based on the opening/closing interactions outlined above, it is not unreasonable to suggest that speech movements are organized minimally across movement phases with movement extent and speed specified for units larger than individual opening and closing actions. The size of the organizational unit is at least on the order of a syllable and perhaps larger encompassing something on the order of a stress group (Fowler, 1983; Sternberg et al., 1988). It can further be suggested that modulation of the relative timing of movement phases is an organizational principle used to differentiate many sounds of the language. As such, serial timing and the associated mechanism is an important concept in speech production but one that has not received much empirical attention. It has been suggested previously that the serial order for speech is a consequence of an underlying rhythm generating mechanism (Kozhevnikov & Chistovich, 1965; Lashley, 1951; Kelso & Tuller, 1984; Gracco, 1988; 1990). For Lashley (1951), the simplest of timing mechanisms are those that control rhythmic activity and he drew parallels between the rhythmic movements of respiration, fish swimming, musical rhythms, and speech (see also Stetson, 1951). For speech, however, the mechanism can not be simple or stereotypic since, as shown in this investigation, the different sounds in the language have different temporal requirements. Moreover, such a rhythmic mechanism most likely reflects a network property (Martin, 1972) as opposed to a property of a localized group of neurons, since dysprosody results from damage to many different regions of the nervous system (e.g.., Kent & Rosenbeck, 1982). More likely, any centrally generated rhythmic mechanism for speech and any other flexible behavior must be modifiable by internal and external inputs (cf. Getting, 1989; Glass & Mackey, 1988; Harris-Warrick, 1988; Rossignol, Lund, & Drew, 1988).

Recent investigations on the discrete effects of mechanical perturbation on sequential speech movements are beginning to identify aspects of the underlying serial timing mechanism . Mechanical perturbations to the lower lip during sequential movement result in an increase or decrease in the movement cycle frequency depending on the phase of the movement the load is applied. (Gracco & Abbs, 1988; 1989; Saltzman, Kay, Rubin, & K.-Shaw, 1991). One interpretation of these results is that a rhythmic mechanism is the foundation for the serial timing of speech movements and that this mechanism is modifiable not

stereotypic (Gracco, 1990; 1991). The present results demonstrating changes in opening and closing phasing with phonetic context reflect another aspect of the serial timing for speech. The consonants acted to differentially modify the ongoing speech rhythm. As such, consonants can be considered to act as perturbations on an underlying vowel-dependent rhythm. Consistent with this speculation is evidence from two apparently contradictory results obtained by Ohala (1975) and Kelso et al. (1985). Searching for the underlying preferred speech movement frequency Ohala (1975) was unable to find an isochronous frequency associated with jaw movements during oral reading. In contrast, Kelso et al. (1985) provided evidence suggesting that jaw movements during reiterant speaking were produced with very little variation around a characteristic frequency. The difference in these two investigations is in the phonetic content used by the different investigators. The reiterant speech contained the syllable "ba" produced with different stress levels whereas Ohala used unconstrained oral reading passages. Sounds of the language may have an inherent frequency (intrinsic timing; Fowler, 1980) which interacts with a central rhythm resulting in a modal frequency with significant dispersion. Similarly, the degree to which a characteristic rhythm or frequency is modulated during speaking is no doubt a result of an interaction of a number of functional considerations such as the articulatory adjustments for the specific phoneme, and other suprasegmental considerations such as stress and rate (Gracco, 1990; 1991). While speculative, further investigations of the serial timing characteristics or rhythm generation for speech (Gracco & Abbs, 1989; Saltzman et al., 1991) combined with computational models to evaluate the potential oscillatory network or serial dynamics (Laboissiere et al., 1991; Saltzman & Munhall, 1989) associated with the temporal patterning for speech are important areas of future research.

Speech production: A functional neuromotor perspective

The observations above suggest that the neuromotor representation for speech is more complicated than can be captured by a single control variable or sensorimotor mechanism. Rather, speech production is organized according to principles that incorporate the sequential nature of the process, the functional character of the production units, and the articulatory details that shape the acoustic output. It appears that speech motor control is organized at a functional level according to sound-producing vocal tract actions (Gracco, 1990; 1991; Kent, Carney, & Severeid, 1974). These functional groupings are stored in memory and map categorically onto the sounds of the language. The apparent constraint on coordinating functionally-related articulators observed in this and other investigations (Gracco, 1988; 1990; Gracco & Löfqvist, 1989; in preparation) suggests that the temporal patterning among articulators is a component of the neuromotor representation. Observable speech movements (or vocal tract area functions) reflect a combination of stored representations with flexible sensorimotor processes interacting to form complex vocal tract patterns from relative simple operations (Gracco & Abbs, 1987; Gracco, 1990; 1991). Stored neuromotor representations and sensorimotor interactions simplify the overall motor control process by minimizing much of the computational load. Important areas of future research are to determine the precise role and contribution of biomechanical properties to the observable patterns, to evaluate the relative strength of articulator interactions, and to identify how articulator movements are modified by linguistic and cognitive considerations. The major contributions to understanding speech motor control and underlying nervous system organization will come from a better understanding of the neural, physical, and cognitive factors that govern this uniquely human behavior.

REFERENCES

Abbs, J. H., & Gracco, V. L. (1984). Control of complex motor gestures: Orofacial muscle responses to load perturbations of the lip during speech. *Journal of Neurophysiology, 51(4)*, 705-723.

Adams, S. G., Weismer, G., & Kent, R. D. (1993). Speaking rate and speech movement velocity profiles. *Journal of Speech and Hearing Research, 36*, 41-54.

Arkebauer, H. J., Hixon, T. J., &, Hardy, J. C. (1967). Peak intraoral air pressures during speech. *Journal of Speech and Hearing Research, 10*, 196-208.

Barlow, S. M., Cole, K. J., &, Abbs, J. H. (1983). A new head-mounted lip-jaw movement transduction system for the study of motor speech disorders. *Journal of Speech and Hearing Research, 26*, 283-288.

Bernstein, N. (1967). *The co-ordination and regulation of movements*. New York: Pergamon Press.

Black, J. W. (1950). The pressure component in the production of consonants. *Journal of Speech and Hearing Disorders, 15*, 207-210.

Browman, C. P., & Goldstein, L. M. (1989). Tiers in articulatory phonology, with some implications for casual speech. In J. Kingston & M. Beckman (Eds.), *Papers in laboratory phonology I: Between the grammar and physics of speech* (pp. 341-376). Cambridge, England: Cambridge University Press.

Browman, C. P., & Goldstein, L. M. (1990). Articulatory gestures as phonological units. *Phonology, 6*:2.

Caruso, A. J., Abbs, J. H., & Gracco, V. L. (1988). Kinematic analysis of multiple movement coordination during speech in stutterers. *Brain, 111*, 439-455.

Chen, M. (1970). Vowel length variation as a function of the voicing of the consonant environment. *Phonetica, 22*, 129-159.

Cooke, J. D. (1980). The organization of simple, skilled movements. In G. E. Stelmach & J. Requin (Eds.), *Tutorials in motor behavior* (pp. 199-212). Amsterdam: North-Holland.

Daniloff, R., & Moll, K. (1968). Coarticulation of lip rounding. *Journal of Speech and Hearing Research, 11*, 707-721.

Denes, P. (1955). Effect of duration on the perception of voicing. *Journal of the Acoustical Society of America, 27*, 761-764.

DeNil, L. & Abbs, J. (1991). Influence of speaking rate on the upper lip, lower lip, and jaw peak velocity sequencing during bilabial closing movements. *Journal of the Acoustical Society of America, 89*, 845-849.

Edwards, J., Beckman, M. E., & Fletcher, J. (1991). The articulatory kinematics of final lengthening. *Journal of the Acoustical Society of America, 89*, 369-382.

Folkins, J. W. (1981). Muscle activity for jaw closing during speech. *Journal of Speech and Hearing Research, 24*, 601-615.

Folkins, J. W, & Abbs, J. H. (1975). Lip and jaw motor control during speech: responses to resistive loading of the jaw. *Journal of Speech and Hearing Research, 18*, 207-220.

Folkins, J. W., & Linville, R. N. (1983). The effects of varying lower-lip displacement on upper-lip movements: Implications for the coordination of speech movements. *Journal of Speech and Hearing Research, 26*, 209-217.

Folkins, J. W., & Canty, J. (1986). Movements of the upper and lower lips during speech: Interactions between lips with the jaw fixed at different position. *Journal of Speech and Hearing Research, 29*, 348-356.

Folkins, J. W., & Brown, C. K. (1987). Upper lip, lower lip, and jaw interactions during speech: Comments on evidence from repetition-to-repetition variability. *Journal of the Acoustical Society of America, 82*, 1919-1924.

Fowler, C. A. (1980). Coarticulation and theories of intrinsic timing. *Journal of Phonetics, 8*, 113-133.

Fowler, C. A. (1983). Converging sources of evidence on spoken and perceived rhythms of speech: Cyclic production of vowels in monosyllabic stress feet. *Journal of Experimental Psychology: General., 112*, 386-412.

Fowler, C. A. (1992). Vowel duration and closure duration in voiced and unvoiced stops: there are no contrast effects here. *Journal of Phonetics, 20*, 143-165.

Fowler, C. A., Gracco, V. L., V.-Bateson, E. A., & Romero, J. (submitted) Global correlates of stress accent in spoken sentences. *Journal of the Acoustical Society of America*.

Fromkin, V. A. (1966). Neuro-muscular specification of linguistic units. *Language and Speech, 9*, 170-199.

Fujimura, O., & Miller, J. (1979). Mandible height and syllable-final tenseness. *Phonetica, 36*, 2630272.

Getting, P. A. (1989). Emerging principles governing the operation of neural networks. *Annual Review of Neuroscience, 12*, 185-204.

Glass, L., & Mackey, M. C. (1988). *From clocks to chaos*. Princeton: Princeton University Press.

Gracco, V. L. (1987). A multilevel control model for speech motor activity. I H. Peters & W. Hulstijn (Eds.), *Speech motor dynamics in stuttering* (pp. 57-76). Wien: Springer-Verlag, .

Gracco, V. L. (1988). Timing factors in the coordination of speech movements. *Journal of Neuroscience, 8*, 4628-4634.

Gracco, V. L. (1990) Characteristics of speech as a motor control system. In G. Hammond (Ed.), *Cerebral control of speech and limb movements* (pp. 3-28). North Holland: Elsevier, .

Gracco, V. L. (1991). Sensorimotor mechanisms in speech motor control. In H. Peters, W. Hultsijn, & C. W. Starkweather (Eds.), *Speech motor control and stuttering* (pp. 53-78).North Holland: Elsevier.

Gracco, V. L., & Abbs, J. H. (1985). Dynamic control of the perioral system during speech: kinematic analyses of autogenic and nonautogenic sensorimotor processes. *Journal of Neurophysiology, 54*, 418-432.

Gracco, V. L., & Abbs, J. H. (1986). Variant and invariant characteristics of speech movements. *Experimental Brain Research, 65*, 156-166.

Gracco, V. L., & Abbs, J. H. (1987). Programming and execution processes of speech motor control. Potential neural correlates. In E. Keller, & M. Gopnick, Eds.), *Motor and sensory processes of language* (pp. 163-202). Hillsdale, NJ: Lawrence Erlbaum.

Gracco, V. L., & Abbs, J. H. (1988). Central patterning of speech movements. *Experimental Brain Research, 71*, 515-526.

Gracco, V. L., & Abbs, J. H. (1989). Sensorimotor characteristics of speech motor sequences. *Experimental Brain Research, 75*, 586-598.

Gracco, V. L.., & Löfqvist (1989). Speech movement coordination: Oral-laryngeal interactions. *Journal of the Acoustical Society of America, 86*, S114.

Harris, K. S., Lysaught, G. F., &, Schvey, M. M. (1965). Some aspects of the production of oral and nasal labial stops. *Language and Speech, 8*, 135-147.

Harris-Warrick, R. M. (1988). Chemical modulation of central pattern generators. I, A. Cohen, S. Rossignol, & S. Grillner (Eds.), *Neural control of rhythmic movements in vertebrates* (pp. 285-332). John Wiley & Sons: New York.

House, A. S., & Fairbanks, G. (1953). The influence of consonant environment upon the secondary acoustic characteristics of vowels. *Journal of the Acoustical Society of America, 25*, 105-113.

Hughes, O., & Abbs, J. (1976). Labial-mandibular coordination in the production of speech: implications for the operation of motor equivalence. *Phonetica, 33*, 199-221.

Kelso, J. A. S. (1986). Pattern formation in speech and limb movements involving many degrees of freedom. In H. Heuer & C. Fromm (Eds.), Generation and modulation of action patterns (pp. 105-128). Berlin: Springer-Verlag.

Kelso, J. A. S., & Tuller, B. (1984). Converging evidence in support of common dynamical principles for speech and movement coordination. *American Journal of Physiology, 15*, R928-R935.

Kelso, J. A. S., Tuller, B., V.-Bateson, E., & Fowler, C. (1984). Functionally specific articulatory cooperation following jaw perturbations during speech: Evidence for coordinative structures. *Journal of Experimental Psychology: Human Perception and Performance, 10*, 812-832.

Kelso, J. A. S., V.-Bateson, E., Saltzman, E. L., & Kay, B. (1985). A qualitative dynamic analysis of reiterant speech production: Phase portraits, kinematics, and dynamic modeling. *Journal of the Acoustical Society of America, 77*, 266-280.

Kent, R. D. (1983). The segmental organization of speech. In P. MacNeilage (Ed.), *The production of speech*. New York: Springer-Verlag.

Kent, R. D., Carney, P. J., & Severeid, L. R. (1974). Velar movement and timing: evaluation of a model for binary control. *Journal of Speech and Hearing Research, 17*, 470-488.

Kent, R. D., & Rosenbek, J. C. (1982). Prosodic disturbances and neurologic lesion. *Brain Language, 15*, 259-291.

Kluender, K. R., Diehl, R. L., & Wright, B. A. (1988). Vowel-length differences before voiced and voiceless consonants: An auditory explanation. *Journal of Phonetics, 16*, 153-169.

Kollia, Gracco & Harris (1992). Functional organization of velar movements following jaw perturbation. *Journal of the Acoustical Society of America, 91(2)*, 2474.

Kozhevnikov, V., & Chistovich, L. (1965). *Speech: Articulation and perception*. Joint Publications Research Service, 30,453; U.S. Department of Commerce.

Kuehn, D. P. & Moll, K. (1976). A cineradiographic study of VC and CV articulatory velocities. *Journal of Phonetics, 4*, 303-320.

Kuehn, D. P., Reich, A. R., & Jordan, J. E. (1980). A cineradiographic study of chin marker positioning: Implications for the strain gauge transduction of jaw movement. *Journal of the Acoustical Society of America, 67*, 1825-1827.

Laboissiere, R., Schwartz, J.-L., & Bailly, G. (1991). Motor control for speech skills: A connectionist approach. In D. S. Touretzky & G. E. Hinton (Eds.), *Proceedings of the 1990 Connectionist Models Summer School* (pp. 319-327). San Mateo, CA: Morgan Kaufmann.

Lashley, K. S. (1951). The problem of serial order in behavior. In L. A.. Jeffress (Ed.), *Cerebral mechanisms in behavior: The Hixon symposium*. New York: Wiley.

Löfqvist, A., & Yoshioka, H. (1981). Interarticulator programming in obstruent production. *Phonetica, 38*, 21-34.

Löfqvist, A., & Yoshioka, H. (1984). Intrasegmental timing: Laryngeal-oral coordination in voiceless consonant production. *Speech Communication, 3*, 279-289.

Lindblom, B. E. F. (1987). Adaptive variability and absolute constancy in speech signals: two themes in the quest for phonetic invariance. *Proceedings of the Eleventh International Congress of Phonetic Sciences, 3*, 9-18.

Lindblom, B. E. F., Lubker, J., & Gay, T. (1979). Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation. *Journal of Phonetics, 7*, 147-161.

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: acoustical measurements. *Word, 20*, 384.

Lubker, J. F., & Parris, P. J. (1970). Simultaneous measurements of intraoral pressure, force of labial contact, and labial electromyographic activity during production of the stop consonant cognates /p/ and /b/. *Journal of the Acoustical Society of America, 47*, 625-633.

Luce, P. A., & Charles-Luce, J. (1985). Contextual effects on vowel duration, closure duration and the consonant/vowel ratio in speech production. *Journal of the Acoustical Society of America, 78*, 1949-1957.

Macchi, M, (1988). Labial articulation patterns associated with segmental features and syllable structure in English. *Phonetica, 45*, 109-121.

Martin, J. G. (1972). Rhythmic (hierarchical) versus serial structure in speech and other behavior. *Psychological Review, 79*, 487-509.

McClean, M. D., Kroll, R. M., & Loftus, N. S. (1990). Kinematic analysis of lip closure in stutterers' fluent speech. *Journal of Speech and Hearing Research, 33*, 755-760.

Moore, C.A., Smith, A., & Ringel, R. L.. (1988). Task-specific organization of activity in human jaw muscles. *Journal of Speech and Hearing Research, 31*, 670-680.

Müller, E. M., & Abbs, J. H. (1979). Strain gage transduction of lip and jaw motion in the midsagittal plane: Refinement of a prototype system. *Journal of the Acoustical Society of America, 65*, 481-486.

Munhall, K., Ostry, D. J., & Parush, A. (1985). Characteristics of velocity profiles of speech movements. *Journal of Experimental Psychology: Human Perception and Performance, 11*, 457-474.

Munhall, K., Löfqvist, A., & Kelso, J. A. S. (in press). Lip-larynx coordination in speech: effects of mechanical perturbations to the lower lip. Journal of the Acoustical Society of America.

Nelson, W. L. (1983). Physical principles for economies of skilled movements. *Biological Cybernetics, 46*, 135-147.

Ohala, J. J. (1975). The temporal regulation of speech. In, G. Fant & M. A. A. Tatham (Eds.), *Auditory analysis and perception of speech* (pp. 431-454). London: Academic.

Ostry, D. J., Keller, E., & Parush, A. (1983), Similarities in the control of the speech articulators and the limbs: kinematics of tongue dorsum movement in speech. *Journal of Experimental Psychology: Human Perception and Performance, 9*, 622-636.

Ostry, D. J., & Munhall, K. G. (1985). Control of rate and duration of speech movements. *Journal of the Acoustical Society of America, 77*, 640-648.

Parush, A., Ostry, D. O., & Munhall, K. G. (1983). A kinematic study of lingual coarticulation in VCV sequences. *Journal of the Acoustical Society of America, 74*, 1115-1125.

Perkell, J. S. (1969). Physiology of speech production. *Research Monograph (53)*. Cambridge, MA: MIT Press.

Rossignol, S., Lund, J. P., & Drew, T. (1988). The role of sensory inputs in regulating patterns of rhythmical movements in higher vertebrates: A comparison between locomotion, Respiration, and mastication. In A. Cohen, S. Rossignol, & S. Grillner (Eds.), *Neural control of rhythmic movements in vertebrates* (pp. 201-284). John Wiley & Sons: New York.

Saltzman, E. L. (1986). Task dynamic coordination of the speech articulators: a preliminary model. In H. Heuer & C. Fromm, (Eds.), *Generation and modulation of action patterns* (pp. 129-144). Berlin: Springer-Verlag.

Saltzman, E. L., & Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology, 1(4)*, 333-382.

Saltzman, E. L., Kay, B., Rubin, P., & Kinsella-Shaw, J. (1991). Dynamics of intergestural timing. *PERILUS, 16*, 47-55.

Sternberg, S, Knoll, R. L., Monsell, S., & Wright, C. E. (1988). Motor programs and hierarchical organization in the control of rapid speech. *Phonetica, 45*, 175-197.

Stetson, R. H. (1951). Motor phonetics: A study of speech movements in action. North Holland Publishing Co.

Stevens, K. N. (1972). On the quantal nature of speech: Evidence from articulatory-acoustic data. In E. E. David & P. B. Denes (Eds.), *Human communication: A unified view* (pp. 51-66). New York: McGraw-Hill.

Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics, 17*, 3-45.

Summers, W. V. (1987). Effects of stress and final-consonant voicing on vowel production: Articulatory and acoustic analyses. *Journal of the Acoustical Society of America, 82*, 847-863.

Sussman, H. M., MacNeilage, P. F., & Hanson, R. J. (1973). Labial and mandibular dynamics during the production of bilabial consonants: Preliminary observations. *Journal of Speech and Hearing Research, 16*, 397-420.

Tatham, M. A. A., & Morton, K. (1968). Some electromyography data towards a model of speech production. University of Essex, Language Centre, *Occasional Papers 1*.

von Nuemann, J. (1958). *The computer and the brain*. New Haven: Yale University Press.

Weismer, G. Assessment of articulatory timing. In, J. Cooper (Ed.), Assessment of Speech and Voice Production: Research and Clinical Applications, (pp. 84-95). NIDCD Monograph.

FOOTNOTE

Journal of Speech and Hearing Research, in press.

BEST COPY AVAILABLE

Haskins Laboratories Status Report on Speech Research
1993, SR-113, 91-94

The Quasi-steady Approximation in Speech Production*

Richard S. McGowan

Because boundary-layer separation is important in determining force between flowing air and solid bodies and separation can be sensitive to unsteadiness, the quasi-steady approximation needs to be examined for the flow-induced oscillations of speech (e.g., phonation and trills). A review of the literature shows that vibratory phenomena, such as phonation and tongue-tip trills, may need to be modeled without the quasi-steady approximation.

The quasi-steady approximation is commonly made in modeling air flow and vibration in speech production. Experimentally, the quasi-steady approximation means that time-varying situations, such as vocal fold oscillation or tongue-tip trills, can be studied with a series of static configurations simulating the air flow in a vocal tract. For mathematical modeling purposes, the quasi-steady approximation means that acceleration terms involving partial derivatives with respect to time in the equations of motion of air can be neglected. This allows the modeling of air flow as a sequence of static flow configurations. Although an inductive term representing the effect of acceleration of air in the constriction is often included in mathematical modeling, other unsteady effects are ignored. When the quasi-steady approximation is made, vorticity and turbulence distributions, important in force and energy balance considerations, are assumed to be unaffected by unsteady air acceleration. In particular, the quasi-steady approximation is applied to boundary-layer separation, which is an important determinant of vorticity distribution, and, hence, of the energy exchange between the air and solid in flow-induced oscillations, such as phonation and tongue-tip trills. However, a review of recent literature (e.g., Bertram & Pedley, 1983; Cancelli & Pedley, 1985; Pedley & Stephanoff, 1985; Sobey, 1983; Sobey, 1985) shows that the quasi-steady approximation can only be used with great care in the fluid mechanics regimes involving boundary-layer separation. Because boundary-layer separation occurs in the vocal tract these recent works bear consideration for understanding speech production. Some of the recent literature that brings the quasi-steady approximation into question will be reviewed here, after some of its relevance to speech production modeling has been discussed.

The fact that characteristic Strouhal numbers are often small in speech has been used to justify the quasi-steady approximation. For periodic motion the Strouhal number is the product of a characteristic frequency and characteristic length scale divided by a characteristic velocity, and it is the coefficient of the time derivative terms in nondimensional versions of the mass and momentum conservation equations. If the Strouhal number is small compared to one, it is presumed that these terms can be neglected, that is, it is possible to make the quasi-steady approximation. (This assumes that other terms are of order one,

This work was supported by NIH grants DC-00121 and DC-00865 to Haskins Laboratories. Thanks to Anders Löfqvist and Philip Rubin for comments.

which is generally the case.) The quasi-steady approximation is often used in aerodynamic considerations for the modeling of phonation, stop release, and fricatives. For phonation, with maximum air speeds in excess of 2000 cm/sec, frequencies on the order of 100 Hz, and length scales at most on the order of 1 cm, producing a Strouhal number less than .05, this may appear to be a valid approximation to use for simplifying a model (Catford, 1977, p. 98). Before reviewing the reasons that this may be faulty, it is necessary to consider boundary-layer separation and its importance for vibratory phenomena in the vocal tract.

When air flows over a solid surface, a layer of air with a high concentration of vorticity is formed next to the solid surface, and this layer is a boundary layer. Boundary-layer separation occurs at places on the surface of a solid where the vorticity of the boundary layer abruptly leaves the region close to the solid boundary and is subsequently convected by the flow. The places where separation occurs are called separation points. (A separation point in two-dimensional modeling represents a line of separation in the third dimension.) Separation occurs where there is a sufficient adverse pressure gradient, as occurs when flow is decelerated (Lighthill, 1963). An adverse pressure gradient occurs when there is flow along solids from regions of low pressure to regions of high pressure, and if the spatial rate of change of pressure is sufficiently large, the boundary layer will separate. For example, the flow from the constriction for an /s/ separates because of the abrupt, adverse pressure change caused by the sudden area expansion after the constriction. During phonation, the flow separates from the folds, thus providing the time-varying flow resistances that are essential for the production of modal voice (Ishizaka & Flanagan, 1972).

The locations of separation points are important for the study of flow induced oscillations in the vocal tract because they help to determine the forces between the air and the tissue. Because vorticity is transported from the boundary layer into the main portion of the flow field at a separation point, there can be a change in pressure head in traversing the region near a separation point. A net force on a blunt object in the direction of flow can result from such a pressure head difference. There is likely to be a separation point near the boundary between the windward and the leeward sides of such an object because of the adverse pressure gradient. (There is an adverse pressure gradient because the flow slows down on the leeward side.) The windward side has a higher pressure head than the leeward side, so there is a higher static pressure on the windward side than on the leeward side. These considerations are important for the energy exchange between moving objects and the air, because energy is the integral of force against distance moved. Thus, the change in boundary-layer separation behavior with the removal of the quasi-steady approximation may have important consequences in modeling flow induced oscillations of speech, such as phonation and tongue-tip trills.

In the last decade the quasi-steady approximation for internal flows (e.g., flows inside the vocal tract), even at very small Strouhal number, has been questioned. The Reynolds number, equal to the product of a characteristic velocity and length scale divided by the kinematic viscosity, has been shown to be an important parameter for questions of unsteadiness. Pedley (1983), in a review article on physiological fluid flows, quotes results on channel flows driven sinusoidally from a side wall with amplitudes of between .28 and .57 of the channel width. For Reynolds numbers of between 300 and 700 based on cross-sectionally averaged, steady fluid particle velocity and channel width, quasi-steady behavior disappears for a Strouhal number of about .008. A vortex wave is observed to travel downstream from the oscillating portion of the wall (Pedley & Stephanoff, 1985; Sobey, 1985), which is not predicted by quasi-steady theory. There is also other behavior that marks the inadequacy of the quasi-steady approximation in unsteady flow. Bertram and Pedley (1983) have shown experimentally that impulsively started flow over an indentation of the channel wall can create separated flow on the lee side of the indentation, with the separation point moving upstream as the steady state is approached. Sobey (1983) has used numerical simulations of unsteady flow in channels with wavy walls to show the moving separation point on the lee slopes of the wavy

walls. For a Reynolds number based on peak velocity and minimum channel half-width of only 75 and a Strouhal number of .01, there are qualitative differences in the separation behavior from that expected from the quasi-steady approximation. For one thing the flow, once separated during acceleration, does not reattach to the walls after deceleration to zero flow. Requiring that separation vorticity disappears when the flow reverses in oscillatory flow, Sobey derived a very restrictive relation between Strouhal number and Reynolds number for the quasi-steady approximation to be valid. The Strouhal number must be less than .2 of the square inverse of the Reynolds number. For a Reynolds number of 100 the Strouhal number would need to be less than .00002 to meet Sobey's criterion for quasi-steady behavior. Thus, a small Strouhal number is not sufficient to ensure quasi-steady behavior.

Sobey (1983) furthermore gives an argument as to why unsteady separation phenomena are different from steady separation phenomena, even at very small Strouhal numbers. In unsteady flow, the fluid particle velocity at a point in space can be considered both a function of time and a function of a time-variable Reynolds number. To obtain the total time derivative of the fluid particle velocity, one needs to include a term that is the derivative of the fluid particle velocity with respect to Reynolds number times the time derivative of the Reynolds number. Because flow conditions are singular near a separation point, the derivative of flow velocity with respect to the Reynolds number in a region of a separation point can be quite large. Thus, the total time derivative does not scale exclusively with Strouhal number near a separation point, but also depends on the Reynolds number.

It is easily seen, using Sobey's criterion, that flow-induced oscillations in speech may not be quasi-steady if flow separation is concerned. Based on maximum velocity of 2000 cm/sec and characteristic dimension of 1 cm, the Reynolds number during phonation is about 13000. The Strouhal number for phonation is .05 based on a 100Hz oscillation frequency. Based on an oscillation frequency of 30 Hz, the tongue-tip trill Strouhal number is 1/3 of that for phonation, and the Reynolds number is in the same range as that for phonation (McGowan, 1992). Thus, both these vibratory phenomena should be considered to be truly unsteady and the quasi-steady assumption seriously questioned.

One possible consequence of unsteadiness may be a mechanism for energy exchange from air to solid during vibration. For instance, for essentially one-degree-of-freedom solid motion during falsetto voice there could be a hysteresis in the separation point position between opening and closing phases of the motion. If the separation point tends to be further forward during the opening phase than during the closing phase, the pressure on the upstream portion of the folds could be greater during the opening than during the closing phase. This mechanism could supplement others including a glottal air induction that depends on vocal fold position (Wegel, 1930) and an inductive loading of the supraglottal vocal tract (Flanagan & Landgraf, 1968; Ishizaka & Flanagan, 1972) in accounting for energy exchange.

In this note, it has been shown that the quasi-steady approximation may not be a valid approximation in speech, particularly when flow separation is involved. Unsteady effects may have to be included to account for some aspects of phonation and tongue-tip trills. For instance, mechanisms proposed for the energy exchange from air to the vocal folds when each vocal fold has essentially one degree of freedom may be supplemented with a moving separation point.

REFERENCES

Bertram, C. D., & Pedley, T. J. (1983). Steady and unsteady separation in an approximately two-dimensional indented channel. *Journal of Fluid Mechanics, 130,* 315-345.

Cancelli, C., & Pedley, T. J. (1985). A separated-flow model for collapsible tube oscillations. *Journal of Fluid Mechanics, 157,* 375-404.

Catford, J. C. (1977). *Fundamental problems in phonetics.* Bloomington: Indiana University Press.

Flanagan, J. L., & Landgraf, L. L. (1968). Self-oscillating source for vocal-tract synthesizers. *IEEE Trans. Audio and Electroacoustics, AU-16.* 57-64.

Lighthill, M. J. (1963). Introduction. Boundary layer theory. In L. Rosenhead (Ed.), *Laminar boundary layers.* Oxford: The Clarendon Press.

McGowan, R. S. (1992). Tongue-tip trills and vocal-tract wall compliance. *Journal of the Acoustical Society of America, 91,* 2903-2910.

Ishizaka, K., & Flanagan, J. L. (1972). Synthesis of voiced sounds form a two-mass model of the vocal cords. *Bell System Technical Journal, 51,* 1233-1269.

Pedley, T. J. (1983). Wave phenomena in physiological flows. *Journal of Applied Mathematics, 32,* 267-287.

Pedley, T. J., & Stephanoff, K. D. (1985). Flow along a channel with a time dependent indentation in one wall: The generation of vorticity waves. *Journal of Fluid Mechanics, 160*, 337-367.

Sobey, I. J. (1983). The occurrence of separation in oscillatory flow. *Journal of Fluid Mechanics, 134*, 247-257.

Sobey I. J. (1985). Observation of waves during oscillatory channel flow. *Journal of Fluid Mechanics, 151*, 395-426.

Wegel, R. L. (1930). Theory of vibration of the larynx. *Journal of the Acoustical Society of America, 1, No. 3, Part 2*, 1-21.

FOOTNOTE

*Journal of the Acoustical Society of America, 94, 3011-3013 (1993).

Haskins Laboratories Status Report on Speech Research
1993, SR-113, 95-106

Implementing a Genetic Algorithm to Recover Task-dynamic Parameters of an Articulatory Speech Synthesizer

Richard S. McGowan

A genetic algorithm was used to recover the dynamics of a model vocal tract from the speech pressure wave that it produced. The algorithm was generally successful in performing the optimization necessary to do this inverse problem: a problem with significance in the psychology, biology and technology of speech. A natural extension of this work is to study speech learning using a classifier system that employs a genetic algorithm.

INTRODUCTION

This paper reports work on the recovery of dynamic parameters used to drive the articulators in a model vocal tract of an articulatory synthesizer from the spectral properties of the model's speech output. A genetic algorithm was used to study this inverse problem. Descriptions of the articulatory synthesizer and its dynamics will be provided as needed in this introduction.

Articulatory synthesizer

The Haskins Laboratories articulatory synthesizer, ASY, (Rubin, Baer, & Mermelstein, 1981) (Figure 1) uses a model vocal tract developed by Mermelstein (1973). The shape of the vocal tract is controlled by the positions of various articulators, examples of which are the jaw, tongue body, tongue tip and the lips. The coordinates of the jaw are specified by the angle of a fixed-length vector, JA, with origin at the condyle and end at the point marked jaw in Figure 1 (Mermelstein, 1973). The position of the tongue body center is specified by a vector relative to the jaw vector, with its origin at the condyle and end at the tongue-body center articulator. This vector has a length, CL, and an angle, CA, relative to the jaw. The tongue tip is specified by a vector with origin on the outline of the tongue body, whose angle, TA, is relative to the jaw and tongue body, and whose length is TL The lips are specified in Cartesian coordinates, with the vertical dimension specifying the dimension of lip closure/opening, and the horizontal specifying protrusion. The upper lip's vertical position, ULV, is in relation to the fixed skull, and the lower lip vertical position, LLV, is in relation to the jaw. The lips are yoked in the horizontal dimension, so this coordinate is specified for both lips as lip horizontal, LH.

The vocal tract formed by the placement of the articulators is assumed to be a variable-area tube supporting acoustic wave propagation. It can be modeled as a linear filter with a rational transfer function, with the poles corresponding to resonances, or formants. The first three formant frequencies provided the data of the speech acoustics used to map back into the dynamics of the vocal tract articulators in the experiments reported here.

Task dynamics and the one-to-many problem

What does it mean to map into a dynamics of the articulators? Why not use an optimization procedure to find the positions of the ASY articulators at any time given the first three formant frequencies? Previous methods for performing the mapping from acoustics to vocal-tract shape have relied on mapping from static frames of acoustic data to static shapes of the vocal tract (e.g., Atal, Chang, Mathews, & Tukey, 1978).

This work was supported by NIH grant DC-01247 to Haskins Laboratories. Thanks goes to Carol Fowler, Philip Rubin and Elliot Saltzman for making very helpful comments on this work.

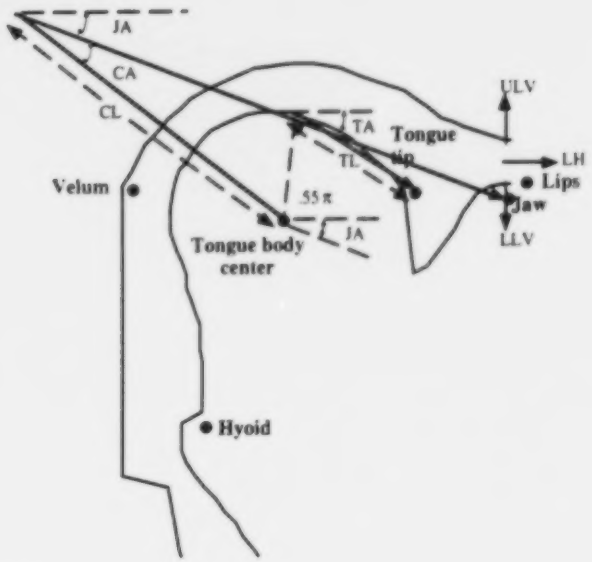


Figure 1. **ASY model vocal tract.**

One problem encountered in these static maps is that of mapping one acoustic signal onto many vocal tract shapes when there is limited acoustic data. In practical terms this indicates that an optimum solution will be nonrobust, so that any perturbation in the data may lead to wildly different results. The constraint that vocal tract shape be continuous in its movement helps to alleviate the one-to-many problem when an impoverished set of acoustic data, such as the first three formant frequencies, is used to map onto a vocal tract shape (e.g., Shirai & Kobayashi, 1986). Alternatively, it has been proposed to map entire trajectories of acoustic data, corresponding to consonant-vowel or vowel-consonant intervals, onto vocal-tract articulator trajectories described as parameterized functions of time (McGowan, 1991). This procedure should be more efficient than mapping several frames of acoustic data onto several vocal tract shapes, optimizing the relation each time, while enforcing continuity.

In experiments on recovering the articulatory movement in the vocal tract, it became apparent that the optimization should be done on a mapping from the acoustic data to task dynamics (Saltzman & Kelso, 1987; Saltzman & Munhall, 1989) instead of the articulator trajectories. The reasons for using task dynamics will be given after its description. Task-dynamics applied to the vocal tract models the coupled motion of the vocal tract articulators in performing speech tasks. For instance, instead of recovering the individual contributions of the jaw and lips in a bilabial closure, the dynamics of the lip aperture reduction would be recovered without particular regard to the individual contributions from the jaw and lip articulators. Figure 2 illustrates the tract variables and lists the articulators associated with each. The focus is the tract variables tongue body constriction location and degree, TBCL and TBCD, tongue tip constriction location and degree TTCL and TTCD, lip aperture, LA, and lip protrusion, LP. TBCL is the location along the vocal tract of the minimum cross-sectional area formed with the tongue body, and TBCD is a measure of the size of that area. Similar interpretations apply to TTCL and TTCD. LA is a measure of the lip opening area, and LP is a measure of the extent to which the vocal tract is lengthened by the lips. The task dynamics of the tract variables is described by a set of uncoupled, second-order, mass-spring systems. Not only are the parameters of natural frequencies and dampings specified, but target positions, and activation intervals (the time during which the task dynamics is active) for target positions other than the default are also specified. Given this set of parameters, the current implementation of task dynamics of the vocal tract describes the formation and releasing of constrictions by the coordinated activity of the component articulators.

ASY ARTICULATOR COORDINATES

($\emptyset_j$; j = 1, 2, ...,n; n=8)

TRACT VARIABLES (Z_i ; i = 1, 2, ...,m; m=6)	LH ($\emptyset_1$)	JA ($\emptyset_2$)	ULV ($\emptyset_3$)	LLV ($\emptyset_4$)	CL ($\emptyset_5$)	CA ($\emptyset_6$)	TL ($\emptyset_7$)	TA ($\emptyset_8$)
LP (Z_1)	●							
LA (Z_2)		●	●	●				
TBCL (Z_3)		●			●	●		
TBCD (Z_4)		●			●	●		
TTCL (Z_5)		●			●	●	●	●
TTCD (Z_6)		●			●	●	●	●

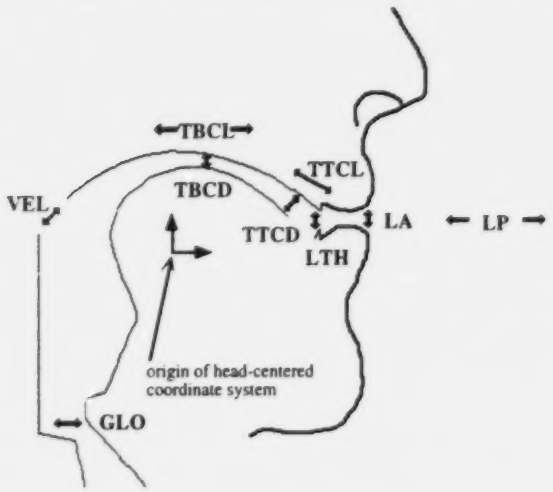


Figure 2. **Tract variables.**

A reason for optimizing for the transformation from acoustics to tract variables rather than for the transformation from acoustics to articulators is that the tract variables specify the acoustically salient features of vocal tract shape, constriction location and degree, more directly than do the articulators. Boë, Perrier, and Bailly (1992) have noted that acoustic output, in terms of formant frequencies, seems to be most sensitive to place and degree of constriction and emphasized that this fact should be used in acoustic-to-articulatory mapping. While it may be true that there is strictly only one articulatory specification for a given vocal tract shape, there may be many and disparate sets of articulatory coordinates that are near by, in the sense of producing similar acoustic output. For instance, an alveolar constriction (the tongue tip close to the ridge behind the upper teeth) can be specified with varying amounts of jaw, tongue body and tongue tip displacement, because these articulators may compensate for each other to attain the prescribed constriction degree and location. While there is not complete compensation throughout the vocal tract, the most

acoustically salient parts of the vocal tract shape, constriction degree and location, can be preserved using compensation. Thus, by mapping to task dynamics, ambiguity caused by articulatory compensation may be removed.

As indicated in Figure 2, tract variables recruit various articulators as their dynamics are instantiated in the vocal tract. The articulators are recruited with various weightings to achieve the target as specified by task dynamics at the tract variable level. Therefore, in the implementation of task dynamics a mapping between the tract variables and articulators, in this case ASY articulators, is specified. While each tract variable appears to act as an independent dynamical system, some may control the same articulator coordinate simultaneously, as the way tongue body constriction degree, TBCD, and lip aperture, LA, both use the jaw angle, JA. Thus, the set of task dynamics equations transformed into the articulatory space becomes a coupled set of equations. Once the task-dynamic parameters are specified, the articulator position trajectories are specified by this set of coupled differential equations. A fourth-order Runge-Kutta routine is used to obtain the articulators' movements and the resulting acoustic output is obtained from ASY.

In the experiments here, and in previous experiments (McGowan, in press), the task dynamics of test utterances resembling /əbæ/ and /ədæ/ were specified in modified gestural scores generated from the linguistic gestural model of Browman and Goldstein (1990). A gestural score is a file containing the task-dynamic specifications for tract variable activation times, stiffnesses, and other information necessary to run the task-dynamic simulation. For each utterance the first three formant frequencies were extracted at 10 ms intervals, thus creating the formant frequency trajectories for the acoustic data. The cost function, or inverse fitness function, was the sum over time of squares of the differences of each of the first three lowest formant frequencies produced by a proposed dynamics and the those of the data.

The genetic algorithm

Since there was no analytic expression between the task-dynamic parameters and the cost function (a numerical extrapolation of nonlinear differential equations was necessary to evaluate the cost function), it was difficult to find the necessary derivatives between the cost function and the task dynamic parameters. Attempting to calculate derivatives numerically would have been very time consuming, because the function evaluations were so complicated. A genetic algorithm was chosen as a nonderivative-based algorithm for optimization of the task-dynamic parameters in gestural scores. A genetic algorithm was not the only possibility for a non-derivative based algorithm, but it had several advantages. These advantages will be explained after the algorithm is described. Genetic optimization procedures are stochastic procedures and there has been some precedence for stochastic procedures in research on the inverse problem. Schroeter, Meyer, and Parthasarthy (1990) have used random access of codebooks in their approach to the inverse problem in speech.

The specific genetic algorithm employed for this study was a modified version of an algorithm described by Goldberg (1989a). In a genetic algorithm, the individuals of a population are assigned randomly chosen parameter sets that are coded into binary strings called "chromosomes", and each is assigned a fitness. The fitness used in this study was the inverse of the sum, taken over 10 ms intervals, of the squares of the difference between the formant frequencies of a proposed solution and those of the speech data, for the first three formant frequencies. Based on fitness, individuals are chosen to breed with others to form a new population of chromosomes. When two individuals mate their chromosomes split at a randomly chosen location with each of two progeny obtaining one part of their chromosome from each parent. The children's fitnesses are evaluated based on their parameter sets as coded in their chromosomes. A small probability of mutation is allowed.

It should be noted that the use of genetic algorithms requires that the parameters for optimization be coded into finite length strings (chromosomes), which limits the range of any parameter and essentially discretizes the parameter space. The degree of discretization of each parameter can be varied to tune the optimization. This was controlled by varying the range of allowed parameters and the number of bits given to code a specific parameter. Because the ranges of starting and ending activation times were both limited to discrete steps and finite range, the potentially infinite set of parameters was made finite.

The coding that was used here was a simple binary code for the real parameters, such as starting activation time. The coded parameters were concatenated to form a complete chromosome. A better way of forming the chromosomes might be to split the binary representations of the parameters so that the most significant bits of all the parame-

ters are grouped together, then the next significant bits are grouped together, and so on. This may be better because of it is more likely that shorter lengths of chromosomes stay together through the mating process and the fitness of any individual depends on how the parameters interact.

There were some particular advantages in employing a genetic algorithm for the optimization procedure in this work. The basic genetic algorithm was relatively easy to implement, and this was an important consideration because of the exploratory nature of this research: not much time would have been spent if things had not worked out. Also, it was easy to watch the optimization procedure evolve, and, thus, to tune the parameters of the genetic algorithm. A very important consideration was that genetic algorithms provide a natural way to bound the ranges of the parameters that are to be optimized, because the parameters are coded in finite length strings. This meant that the task-dynamic model would not be driven beyond certain bounds and numerical overflow during optimization could be avoided. Another important factor was a genetic algorithm's ability to incorporate an on-off gene. For example, the optimal solution may or may not involve the activation of the tongue tip (TTCL and TTCD). It was easy to include a bit in each individual's chromosome that determines whether the tongue-tip was to be activated or not. If the tongue tip was to be activated, then the part of the chromosome containing tongue-tip task parameters could be used, and if not, then that part of the chromosome would be ignored. Finally, the genetic algorithm can be used for purposes other than straight-forward optimization. Speech learning is an obvious extension of the work presented here, so the fact that genetic algorithms can be used in classifier systems is an important consideration.

The power of the genetic algorithm comes from what John Holland, the originator of genetic algorithms, called *implicit parallelism*. Although the population of chromosomes is finite, say number N, the number of patterns of chromosomes, called schemata, being processed is on the order of N^3 (Goldberg, 1989, pp. 40-41). This implicit parallelism, as well as the probability of mutation, makes the algorithm much less likely to become stuck in local maxima. Also, the implicit parallel processing property makes this algorithm more efficient than exhaustive search of the parameter space. Parallel processing in the usual sense of using many processors is also possible. Given that children chromosomes have been produced as the result of mating a generation of individuals, the children's fitnesses can be evaluated on physically distinct processors. This capability was not used in this work.

Significance

Recovery of articulation by optimization is reminiscent of the analysis by synthesis model of speech perception (Stevens, 1960; Stevens & Halle, 1967; Stevens & House, 1970). In analysis by synthesis, as originally conceived, the content of a speech signal (the data) is recovered when hypothesized phonetic segments and features are used to provide instructions to an articulatory mechanism which, in turn, produces speech to be compared with the data. When the comparison is good enough, the hypothesized phonetic segment and features are the perceived phonetic segments and features. Rival theories of speech perception also take recovery of the vocal tract articulatory movement or the intended gesture over time as central to perceiving the spoken message (e.g., Liberman & Mattingly, 1985). Whatever the psychological importance of the articulatory movement or the motor system for speech perception in adults, children develop their speech by listening and by trying to be understood, and thus this learning process may be considered an analysis by synthesis over an extended time scale. Also, because task-dynamics and articulatory movement are slow relative to the variations in the acoustic pressure wave, there are potential technological advantages in recovering vocal-tract movement from the speech signal to achieve bit-rate reduction for transmission (Schroeter & Sondhi, 1992; Sondhi, 1990). In the work presented here, selected task-dynamic parameters were recovered, without regard to phonetic categories: a model of speech perception is not offered. However, there are future directions that this work might take for speech recognition and for modeling speech development.

Extensions to previous work

In previous work using the methods described above, the focus has been on recovering articulatory movement rather than task dynamics (McGowan, in press). Based on limited testing, the method seems to be successful at doing this. While this method may help solve the one-to-many problem in the transformation from acoustics to articulation, it was not known whether the acoustics maps onto the task-dynamics of the tract variables uniquely. In the previous work, the task-

dynamic parameters were either assumed to be on or they were assumed to be off before the optimization was run. The only task parameters allowed to be activated were the ones known to have synthesized the speech data. It may have been possible to obtain nearly identical articulatory movement from very different task dynamics. That is, there may be a one-to-many problem in mapping from articulation to task dynamics, or from acoustics to task dynamics. Thus, for the present study, on-off genes were added so that selected tract variables could either be activated or inactivated. This allowed the optimization itself to select which tract variables to use.

PROCEDURE

Constraining relations between the parameters were used to keep the number of unknown parameters to a minimum. It was assumed that all movements were critically damped, and that the activation intervals were equal to the period based on the natural frequency parameter. Also, the movement periods, and hence the activation intervals were assumed to be at least 100 ms long to avoid movements that were too stiff for the task-dynamic simulation. With these constraints, the unknowns for a given tract-variable activation were the beginning and ending activation times and the target position. There could have been more than one activation of any tract variable, and these could have overlapped in time. Also, the actions of the certain tract variables were grouped according to common constriction goals, so that they would have identical activation intervals. The tongue body tract variables TBCL and TBCD were in one such group, the tongue-tip variables TTCL and TTCD were in another group, and the lip variables LP and LA were in a group. These groups are known as constriction or task spaces.

Gestural scores were designed to produce articulatory movement for utterances that resembled /əbæ/ and /ədæ/, within the constraints mentioned above. The trajectories of the tract variables resulting from the scores for /əbæ/ and /ədæ/ are illustrated in Figures 3 and 4. The first utterance involves moving from neutral position into a bilabial closure. Subsequently, as the lips open, the tongue lowers for the final vowel. The second utterance involves the tongue making an alveolar closure, and then the release of that closure with a lowering of the tongue body for the final vowel. The heights of the hatched boxes for activation denote the target positions. In recovering the task-dynamic parameters, some parts of the gestural score were taken as known and fixed. In the case of /əbæ/, for the initial movement to lip closure, involving the tract variables LA and LP, the activation interval and targets were taken as known. However, the subsequent activation interval for LA and LP was taken as unknown, as was the target position for LA. Only the target position for LP was assumed known in its second activation. Tongue body movement was taken as unknown, so that the activation interval and the target positions for TBCL and TBCD were varied for the optimum fit, although it was assumed that TBCL and TBCD were activated for at least 100 ms during the utterance. The fact that TTCL and TTCD were not activated for /əbæ/ was assumed to be unknown. The test was to see whether the genetic algorithm would leave these tract variables deactivated in its optimal solution. This can be compared with /ədæ/ where there was activation both of the tongue-tip tract variables, TTCL and TTCD, and the tongue body tract variables, TBCL and TBCD. The lip movement for the final vowel, which occurs after the tongue-tip closure, was taken as completely known. The parameters for the tongue tip were taken as unknown and the algorithm was allowed to turn them on or off. The parameters of the tongue-body tract variables, TBCL and TBCD, in the transition to final vowel were taken as unknown, but they were presumed activated for at least 100 ms some time during the utterance as for /əbæ/.

Given the synthetic acoustic data, a genetic algorithm (Goldberg, 1989a) was used to recover task-dynamic parameters when random noise with a flat distribution between -10 and +10 Hz was added to each formant frequency data at each time frame. The noise was used to test robustness of the procedure. For each test an initial population of chromosomes of 60 individuals was generated using a random number generator. Their fitnesses were evaluated by decoding the chromosomes of each individual into the task dynamic parameters of a gestural score, which enabled the task-dynamic simulation to drive the articulatory synthesizer, ASY, which, in turn produced formant trajectories. The coding was done so that task-dynamic parameters were placed contiguously coded into binary strings with all parameters in each task space grouped together. These parameters included the beginning of activation, the activation interval, and the targets of the tract variables within a given task space. Table 1 indicates the range of the target values for each tract variable, the number of bits used in the coding of that parameter into the chromosomes, and the resulting resolution of that target.

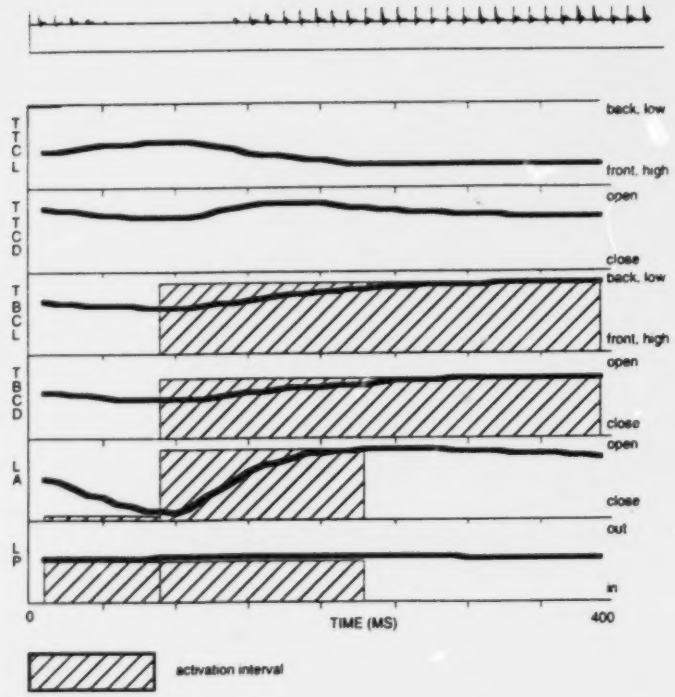


Figure 3. **Original gestural score for /əbæ/.**

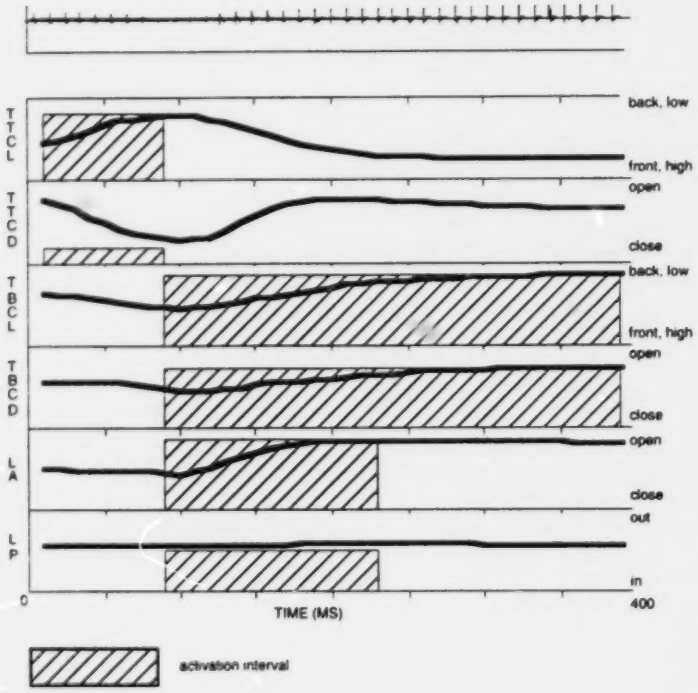


Figure 4. **Original gestural score for /ədæ/.**

Table 1. *Target value specifications.*

Tract Variable	Maximum/Minimum Target Value	Number of Bits in Chromosome	Resolution
LA	1.80/-.3 cm	6	.033 cm
TBCL	3.16/.51 rad	6	.042 rad
TBCD /əbæ/	1.80/-.30 cm	6	.033 cm
TBCD /ədæ/	1.63/-.13 cm		.028 cm
TTCL	1.16/.40 rad	6	.012 rad
TTCD	2.15/-.65 cm	6	.044 cm

Beginning and ending times for the activation intervals were resolved within the data frame rate of 10 ms. Recall that the fitness of an individual was the inverse of the sum of squares of the differences of each of the lowest three formant frequencies produced by the individual and that of the data in 10 ms steps. Thirty pairs of individuals were chosen with probability proportional to each of their fitnesses. There was a .6 chance of each of these pairs to mate, where mating consisted of randomly selecting a cut point for the pair, and then exchanging substrings on each side of the cut point to form two children strings. There was also a .001 chance of mutation at a single position in any of the children. If mating did occur, the fitness of the children strings would then be evaluated. In a variation of the standard genetic algorithm, the best individual of a given generation was always retained into the succeeding generation.

The choice of pairs and possible mating was allowed to go for 60 generations. At the end of 60 generations the individual with the greatest fitness had its string decoded into task-dynamic parameters of a gestural score, which were saved for later comparison. This procedure was repeated 8 times for a new initial random population of 60 individuals. The saved gestural scores were used to drive ASY, and the gestural score and articulator trajectories of the best individual of each run were compared to those that generated the original acoustic data.

RESULTS

The fittest individual of the 8 runs for each utterance is termed the optimal solution, and the individuals having the highest fitness in each of the other runs are termed suboptimal solutions. Figures 5 and 6 show the optimal recovered gestural scores for the utterances /əbæ/ and /ədæ/, respectively. A comparison between Figures 3 and 5 shows that the gestural score for /əbæ/ was well recovered. There were some discrepancies, because the activation interval for the tongue body (TBCL and TBCD) was delayed, while the activation interval for (LA and LP) was slightly delayed, as well as shortened. The recovered paths of the tract variables matched closely those of the original. It is significant here that the tongue tip variables (TTCL and TTCD) were not activated in the optimal solution even though they showed movement because of their dependence on the tongue body. Thus, the procedure was able to correctly decide whether the tongue tip was under active control or not in this instance. The optimal recovery of /ədæ/ seems to be very good (Figures 4 and 6), except that the activation interval for the tongue body (TBCL and TBCD) started late and was too short. In this utterance, not only did the genetic algorithm have the opportunity to shut off the tongue tip (TTCL and TTCD), but there was a sequence of gestures involving the tongue tip and tongue body for which dynamic parameters were unknown. In this instance, the procedure found that the tongue tip was activated.

Quantitative measures of the goodness of fit are indicated by the mean square error in the formant frequency fits. In the optimal fit for /əbæ/ the mean square error was 1099, which corresponds to an average error of 19 Hz per formant per frame of data. For /ədæ/ the mean square error for the optimal fit was 4444, which corresponds to an average error of 38 Hz per formant per frame of data. Using the recovered gestural scores it was possible to generate the articulatory trajectories themselves, such as that for jaw angle, JA (see Figure 1). Tables 2 and 3 show that both the RMS error and maximum error in the optimal recovery of important articulatory trajectories are small in absolute terms.

Perhaps as important as evaluating the performance of the best, or optimal individual, is to evaluate the performances of other, suboptimal individuals. For /əbæ/, the second fittest suboptimal individual was a gestural score that activated the tongue tip (Figure 7). As can be seen, the activation interval of the tongue tip was just over the minimum 100 ms, and the targets were close to neutral position, which was the direction of movement for TTCL and TTCD during that interval anyway. Therefore, this activation had little effect on the tongue tip. The mean square error of this fit was 2004, which corresponds to an average error of 26 Hz per formant per time frame of data The resulting error in the articulation is shown in Table 2, where the absolute errors were almost as small as those for the optimal solution.

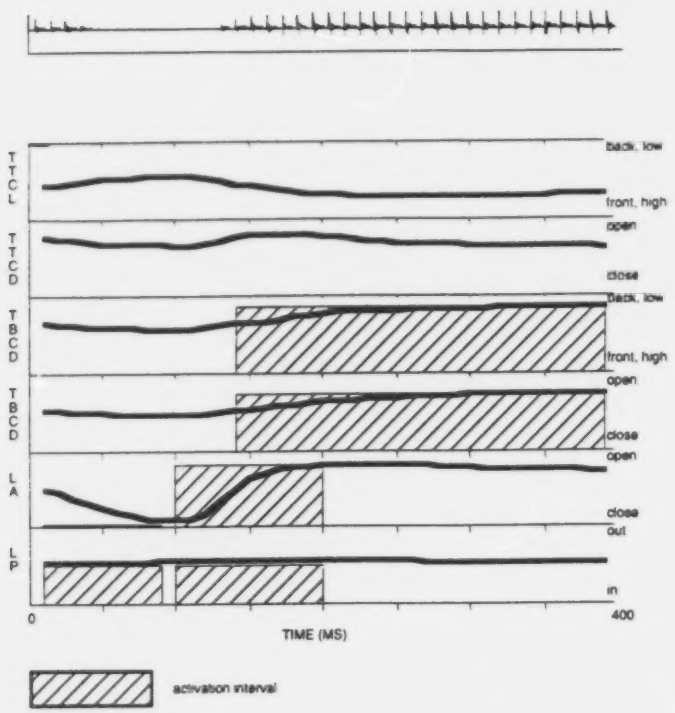


Figure 5. **Optimal gestural score for /əbæ/.**

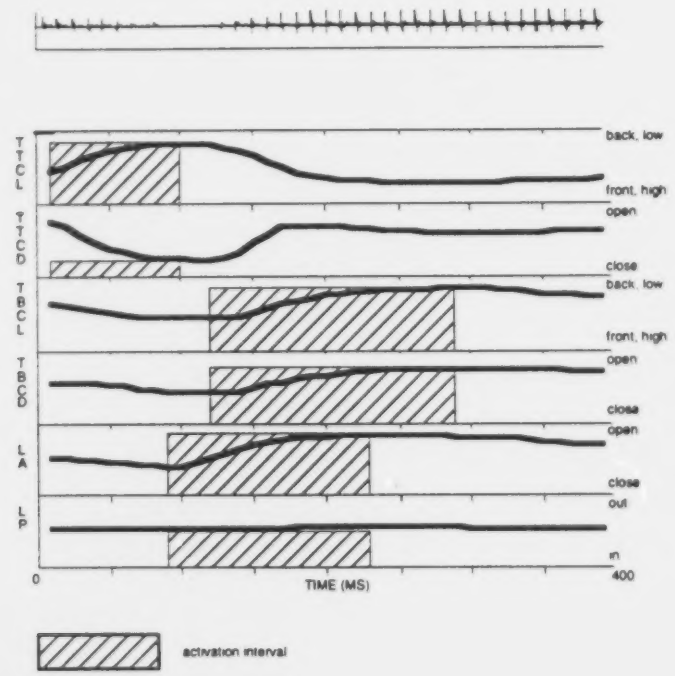


Figure 6. **Optimal gestural score for /ədæ/.**

Table 2. *Comparison of original and recovered trajectories for utterance /əbæ/.*

Articulator coordinate	RMS difference optimal solution	RMS difference suboptimal solution (example)	Maximum absolute difference optimal solution	Maximum absolute difference suboptimal solution (example)
CL	0.0083 cm	0.0067 cm	0.079 cm	0.18 cm
CA	0.00023 rad	0.0019 rad	0.0050 rad	0.019 rad
JA	0.00043 rad	0.0032 rad	0.0070 rad	0.029 rad

Table 3. *Comparison of original and recovered trajectories for utterance /ədæ/.*

Articulator coordinate	RMS difference optimal solution	RMS difference suboptimal solution (example)	Maximum absolute difference optimal solution	Maximum absolute difference suboptimal solution (example)
CL	0.023 cm	0.10 cm	0.26 cm	1.26 cm
CA	0.0037 rad	0.0040 rad	0.035 rad	0.051 rad
TL	0.012 cm	0.098 cm	0.23 cm	1.12 cm
TA	0.0034 rad	0.0035 rad	0.067 rad	0.42 rad
JA	0.0033 rad	0.0049 rad	0.043 rad	0.059 rad

For the utterance /ədæ/, a relatively common feature of the suboptimal solutions was resequencing the tongue tip and tongue body activation intervals. Because these tract variables are coupled through the anatomy of the model articulators, these tract variables could compensate for one another, with the tongue body activated to make the initial alveolar constriction for the /d/ and the tongue tip activated to make the vowel /æ/. There were 4 suboptimal solutions that resequenced the tongue tip and tongue body activations. The fittest of these individuals (fourth fittest overall) is shown in Figure 8. This fit had a mean square error in the formant fit equal to 21584, corresponding to an average error of 85 Hz per formant per time frame of data. It is apparent from Figure 8 that the tongue body was activated to make the initial tongue-tip closure, and that the tongue tip was activated to make the body move down and back for the vowel /æ/. Table 3 shows that the articulation with this suboptimal solution is not nearly as close to the original utterance as the optimal solution: there are large maximum errors in the vector lengths CL and TL.

CONCLUSION

These results indicate that the genetic algorithm is useful for recovering task dynamics. However, the suboptimal solutions show that the algorithm can be driven into locally optimal solutions with incorrect tract variable activations. In the case of /əbæ/ a suboptimal solution produced articulatory paths close to the original utterances despite activating the tongue tip. This indicated that there is a one-to-many situation in the mapping from articulation or acoustics into the task dynamics of the tongue tip. In the case of /ədæ/, the example of a suboptimal solution did not produce an utterance all that close to the original, but the 60 individuals after 60 generations in this particular run were nearly identical, suggesting that the algorithm had found a locally optimal solution far from the optimal solution. In general, it is important to run the genetic algorithm several times to ensure that the proposed suboptimal solution is not trapped in a locally optimal region far from the optimal region.

There are some improvements in the algorithm that can be tried so that local optima might be avoided. One is to code the parameters in the chromosomes so that the most arithmetically significant bits of all the parameters are close by, and the next significant bits are all close by, and so on. In this way the likelihood of breaking apart chromosomes with roughly good combinations of target positions and activation intervals might be avoided. Another improvement would be to run several populations in parallel, and then take the best from each to run again (Goldberg, 1989b).

From the experience gained with optimization, a classifier system based on a genetic algorithm could be used to successfully learn dynamic parameters. Extending the research brings a host of related questions. What happens if there is a mismatch between the synthesizer used for the classifier and the vocal tract that produced the data, as would be the case if the speech data was natural speech? How would the fitness function be affected? For a machine to repeat or learn human speech with an articulatory synthesizer, it does not matter whether the correct dynamics is recovered, as long as the articulatory synthesizer has enough degrees of freedom to produce a very good optimal match with the human acoustic data.

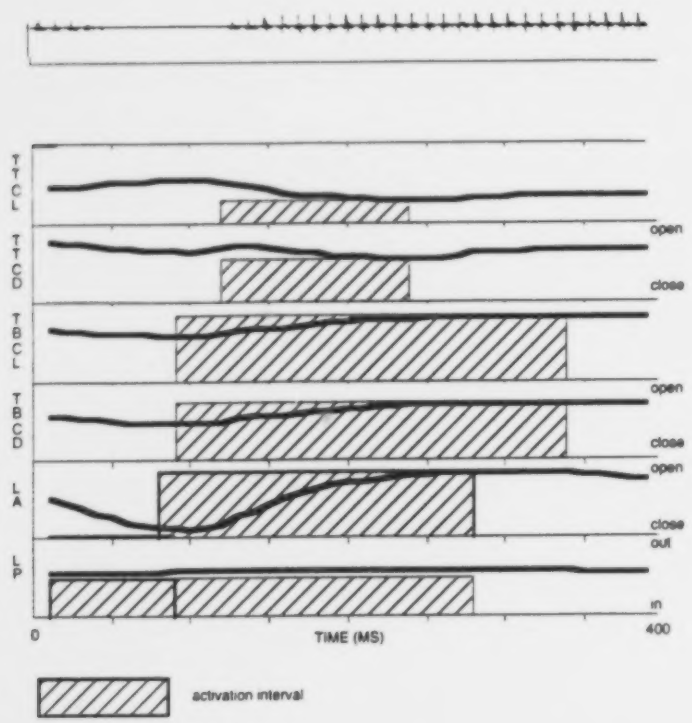


Figure 7. **Suboptimal gestural score for /əbæ/.**

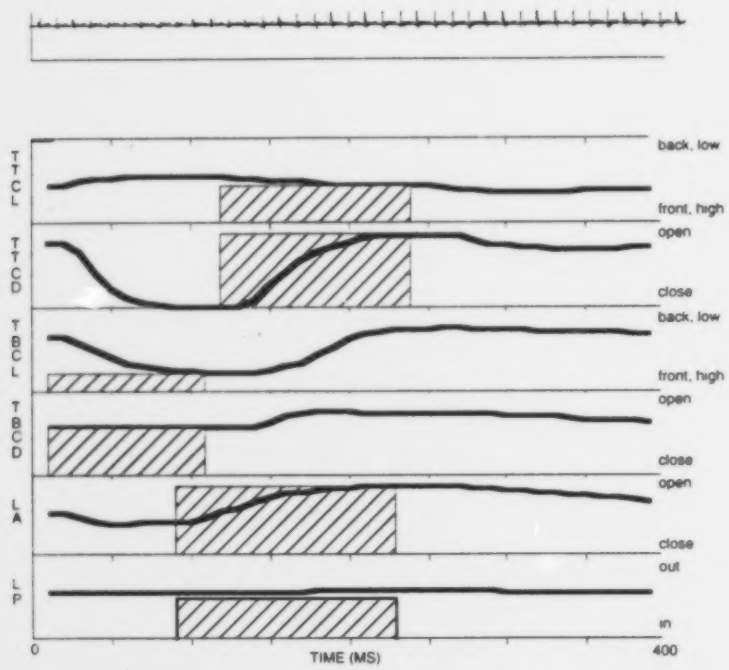


Figure 8. **Suboptimal gestural score for /ədæ/.**

A fixed fitness function evaluated in the acoustic parameter space, analogous to the simple one used here will suffice. If the goal is to model human speech development and the resulting speech perception and production, then it is important to ask whether children learn speech by matching acoustic output. Do children recover the movement of adults' articulators or their dynamics? It is plausible that children neither recover adult articulation, nor match acoustic output. What does this mean for the simulation of human learning? What is the fitness function to use? Children, somehow, recover the adults' meanings and understand adults according to their own needs and desires. They adapt to adult's speaking with the added plus of having a speech apparatus that they can use to be understood (e.g. to get adults to do what they want). A child could construct a dynamics based on his own vocal apparatus and associate others speech sounds with their own dynamics, even when it is impossible to match another's speech sounds exactly, because he can be understood using these movements. Thus, it is not necessary that the child recover adults vocal movements by an analysis by synthesis with a comparator. He knows which of his own movements would produce like meaning. The fitness function would not be based on acoustic match with adults' utterances exactly, but one that involves both meaning and associations with the child's vocal apparatus. The fitness function, in the case of development, would be a "moving target" depending on the state of the adult listeners and speakers, as well as the child's state.

REFERENCES

Atal, B. S ., Chang, J. J., Mathews, M. V., & Tukey, J. W. (1978). Inversion of articulatory-to-acoustic transformation in the vocal-tract by a computer sorting technique. *Journal of the Acoustical Society of America, 63*, 1535-55

Boë, L-J., Perrier, P., & Bailly, G. (1992). The geometric vocal tract variables controlled for vowel production: proposals for constraining acoustic-to-articulatory inversion. *Journal of Phonetics, 20*, 27-38.

Browman, C. P., & Goldstein, L. (1990). Gestural specification using dynamically-defined articulator structures. *Journal of Phonetics, 18*, 299-320.

Goldberg, D. E. (1989a). *Genetic algorithms in search, optimization, and machine learning*. Reading MA: Addison-Wesley Publishing Company, Inc.

Goldberg, D. E. (1989b). Sizing populations for serial and parallel genetic algorithms. In J. D. Schaffer (Ed.), *Proceedings of the Third International Conference on Genetic Algorithms* (pp. 70-9). San Mateo, California: Morgan Kaufman Publishers, Inc.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revisited. *Cognition, 21*, 1-36.

McGowan, R. S. (1991). Recovering tube kinematics using time-varying acoustic information. In *Proceedings of the 12th International Congress of Phonetic Sciences, August 19-24, Aix-en-Provence* (Volume 4, pp. 486-89). Aix-en-Provence: Universite de Provence, Service des Publications.

McGowan, R. S. (in press). Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary model tests. *Speech Communication*.

Mermelstein P. (1973). Articulatory model for the study of speech production. *Journal of the Acoustical Society of America, 53*, 1070-82.

Rubin, P., Baer, T., & Mermelstein, P. (1981). An articulatory synthesizer for perceptual research. *Journal of the Acoustical Society of America, 70*, 321-28.

Saltzman, E. L., & Kelso, J. A. S. (1987). Skilled actions: A task-dynamic approach. *Psychological Review, 94*, 84-106.

Saltzman, E. L., & Munhall, K. G. (1989). A dynamic approach to gestural patterning in speech production. *Ecological Psychology, 14*, 333-82.

Schroeter, J., Meyer, P., & Parthasarthy, S. (1990). Evaluation of improved articulatory codebooks and codebook access distance measures. *IEEE Proceedings, ICASSP 90*, Albuquerque, New Mexico.

Schroeter, J., & Sondhi, M. M. (1992). Speech coding based on physiological models of speech production. In S. Furui & M. M. Sondhi (Eds.), *Advances in speech signal processing* (pp. 231-268). New York: Marcel Dekker, Inc.

Shirai, K., & Kobayashi, T. (1986). Estimating articulatory motion from speech wave. *Speech Communication, 5*, 159-70.

Sondhi, M. M. (1990). Models of speech production for speech analysis and synthesis. *Journal of the Acoustical Society of America, 87*, S14 (A)

Stevens, K. N. (1960). Toward a model for speech recognition. *Journal of the Acoustical Society of America, 32*, 47-55.

Stevens, K. N., & Halle, M. (1967). Remarks on analysis by synthesis and distinctive features. In W. Wathen-Dunn (Ed.), *Models for the perception of speech and visual form*. MIT Press: Cambridge, Massachusetts.

Stevens, K. N., & House, A. S. (1970). Speech perception. In J. Tobias (Ed.), *Foundations of modern auditory theory, Volume II* (pp. 3-63).

Haskins Laboratories Status Report on Speech Research
1993, SR-113, 107-130

An MRI-based Study of Pharyngeal Volume Contrasts in Akan

Mark K. Tiede

Characteristic differences in pharyngeal volume between Akan +/-Advanced Tongue Root (ATR) vowel pairs have been investigated using Magnetic Resonance Imaging (MRI) techniques, and compared to the similar Tense/Lax vowel distinction in English. Two subjects were scanned during steady-state phonation of three pairs of contrasting vowels. Analysis of the resulting images shows that it is the overall difference in pharyngeal volume that is relevant to the Akan vowel contrast, not just the tongue root advancement and laryngeal lowering previously reported from xray studies. The data also show that the ATR contrast is articulatorily distinct from the English Tense/Lax contrast.

INTRODUCTION

Several West African languages have a phonological process of Vowel Harmony that splits the distribution of their vowels into two congruent and mutually exclusive sets. The Akan language, a member of the Niger-Congo Kwa group (Stewart 1971) spoken in Ghana, is typical of this pattern. The vowels of Akan may be grouped as follows:[1]

(1)

Set 1		Set 2	
i	u	ɪ	ʊ
e	o	ɛ	ɔ
a			

In general vowels in an Akan word harmonize; that is all vowels are drawn exclusively from one set or the other. (The low mid vowel /a/ may occur in a word with vowels from either set.[2]) Affixes have two forms for compatibility with stems having type 1 or type 2 vowels. For example (from Dolphyne 1988:15):

(2)

mI	+	di	-->	midi	"I eat"
[1st Sg]		[eat]			
O	+	di	-->	odi	"he eats"
[3rd Sg]		[eat]			
mI	+	dɪ	-->	mɪdɪ	"I am called"
[1st Sg]		[be called]			
O	+	dɪ	-->	ɔdɪ	"he is called"
[3rd Sg]		[be called]			

Drawing on a cineradiographic study of the related language Igbo by Ladefoged (1964), Stewart (1967) proposed an analysis of Akan in which corresponding vowels from the two harmony groups are distinguished articulatorily by differences in tongue root position. Cineradiographic studies of Akan undertaken by Lindau (1975, 1979) confirmed the primacy of tongue root position in the harmony mechanism, but also showed correlated variation in larynx height: advanced tongue root was combined with lowered larynx, and retracted root with raised larynx. In reporting this work Lindau suggested that the relevant contrast was not just the relative positions of these articulators but rather the overall difference in pharyngeal volume produced by their cooperative positioning, and proposed the feature "Expanded" to describe it. A contrast based on differences in pharyngeal volume

I am grateful to the many people who assisted me in this project, particularly Cathe Browman, Alice Faber, Louis Goldstein, Carol Gracco, Robin Greene, Kathy Harris, Pat Nye, Alfred Opoku, Elliot Saltzman, Doug Whalen, and all those who agreed to be magnetized for Science. Goofs and garbles remain of course my responsibility. This work was supported by NIH grant DC-00121.

suggests that vowels from the two groups may also differ in the left-to-right or lateral dimension of the pharynx, assuming that this is subject to voluntary control, but studies based on cineradiography are inherently unable to explore this possibility, since the technique collapses all lateral information into a flat (sagittal) image. The purpose of the study discussed here was to investigate the predicted difference in pharyngeal volume using Magnetic Resonance Imaging, which is not subject to the same limitation.

The magnetic resonance technique has only recently become a viable imaging alternative. Developed primarily for medical diagnostic purposes, MRI exploits the behavior of hydrogen nuclei in a magnetic field to construct an image correlated with the concentration of hydrogen in the scanned tissue.[3] Because different tissue types have differing hydrogen densities and bondings, MR images provide soft tissue definition over a range inaccessible to xray techniques. Two additional advantages make MRI an especially attractive imaging modality: there are currently no known health risks for the subject associated with the technique, and because the imaging plane may be reoriented without moving the subject, three dimensional data collection is possible.

Despite these advantages the MR technique has substantial drawbacks in its potential for phonetic research, chief of which is the tradeoff between imaging time and resulting image quality. Although image acquisition rates continue to drop as the technology evolves, they are currently still too slow by an order of magnitude for capturing dynamic speech. In addition, three dimensional scanning requires multiple passes through the same volume, further increasing acquisition time. This limits the current usefulness of MRI in phonetics to studies involving static vocal tract shapes and sustainable patterns of phonation.

The current study was able to proceed under these constraints. Although vowels sustained for many times their normal speaking duration represent an admittedly artificial source of data, all of the vowels examined (except English lax vowels) can occur in open syllables, and are therefore artificial only in duration, and not in syllable structure. Furthermore, sung vowels of constant pitch and quality and extended duration occur in the musical traditions of Ghanaian and American cultures.[4] There is also precedent for use of the MRI technique applied to measurement of static vowel shapes in the work of Baer, Gore, Gracco, & Nye (in press), who successfully used it to obtain vocal tract area functions of four point vowels, and in similar work by Lakshminarayanan, Lee, and McCutcheon (1991).

The English Tense/Lax distinction eludes precise articulatory description, but is similar in many ways to the Akan contrast: tense vowels are generally articulated with an advanced tongue root, and sometimes a lowered larynx. English vowels comparable to those used in Akan may be grouped as in (1) above:

(3)	Tense		Lax	
	i	u	ɪ	ʊ
	e	o	ɛ	ɔ

Previous attempts to identify the English distinction with the same mechanism used in the Akan ATR contrast include a proposal by Halle and Stevens (1969), and a cineradiographic study by Perkell (1971). However a more extensive cineradiographic study conducted by Ladefoged *et al.* (1972) showed that tongueroot advancement is just one of several complementary strategies used to implement the Tense/Lax contrast, and is not used consistently by all speakers. Similarly electromyographic data reported by Raphael and Bell-Berti (1975) showed differences between subjects in patterns of muscle tension used to distinguish between tense and lax vowels. Factor analysis of xray-derived tongue shapes by Harshman, Ladefoged, and Goldstein (1977) showed that tongue position for English tense/lax pairs can be predicted very completely by reference to just two parameters along which each of the tongue root and tongue dorsum positions covary, whereas a similar analysis of Akan by Jackson (1988) found three parameters necessary for tongue shape specification. While it is probably the case that different mechanisms are involved in each language, they are similar enough to make direct comparison feasible and interesting, and so they have been treated in parallel in the current study.

Method

In this experiment the contrast in cross-sectional area was examined at adjacent levels through the pharynx for corresponding expanded and constricted vowels (i.e. +/-ATR; Tense/Lax).[5] The experiment involved two subjects, both male, in their early thirties. Subject AO is a native speaker of the Asante dialect of Akan; subject MT is a native speaker of Midwestern American English. The vowels selected for comparison were /i : ɪ/, /e : ɛ/, and /u : ʊ/, chosen because of their reasonable similarity across the two languages.

Prior to the experiment a target stimulus tape was prepared for each subject by twice recording each of the target vowels in a characteristic word, extracting the vowel portion using digital editing techniques, then recording it as a continuous utterance by concatenation. The following target words were used:

(4) Vowel		Akan	English	
[i]	pi	"many"	hid	"heed"
[ɪ]	fɪ	"to vomit"	hɪd	"hid"
[e]	ʃɥe	"empty"	heɪd	"hayed"
[ɛ]	ɔʃɥɛ	"he looked at"	hɛd	"head"
[u]	bu	"to break"	hud	"who'd"
[ʊ]	bʊ	"to be drunk"	hʊd	"hood"

These words were also recorded for acoustic analysis:

[o]	ako	"parrot"	hod	"hoed"
[ɔ:ɑ]	kɔɔ	"red" hɑd		"hawed"
[a:æ]	daa	"everyday"	hæd	"had"

The experiment was performed on a General Electric Signa machine installed at the Yale New Haven Hospital. The Signa system consists of a toroidal superconducting electromagnet developing a 1.5 Tesla flux density, placed in a scanning room designed to minimized interference from external electromagnetic noise. The magnet is controlled from an operator's console outside the scanning room, and an attached computer is used for image reconstruction, collation, and storage.

After divesting themselves of all ferrous material subjects were fitted with an earphone, microphone, and neck RF transceiver imaging coil, and positioned on their backs inside the bore of the magnet. The earphone was the terminus of a length of plastic tubing connected to a small speaker placed as far away from the magnet as possible (because the strength of the magnetic field tended to overwhelm the speaker coil). The speaker was driven by an amplifier outside the scanning room, and was used to communicate with the subject and to play the prerecorded target stimuli during scanning. The microphone was used to record subject phonation immediately prior to and following scanning; while phonation during scanning was also recorded it was unusable for analysis because of the intensity of the noise produced by the machine.

Prior to each run the experimenter verified the target vowel by reminding the subject of the characteristic word containing it. Playback of the target vowel was then initiated through the earphone, and scanning begun shortly after the subject began phonating. Subjects were instructed to produce the target vowel with steady pitch and uniform quality, to take shallow breaths while retaining vocal tract configuration, and to refrain from head movement and swallowing as best they could.

Two scans were made for each vowel. The first took approximately two and a half minutes to complete and produced eight adjacent sagittal images (bisecting face) at 3mm intervals (see Table 1 for imaging parameters used). The second scan took approximately three minutes; it produced 28 adjacent images at 5mm intervals in the axial orientation perpendicular to the pharynx (see Table 2).

Table 1. *Sagittal imaging parameters.*

Subject AO (Akan)

256 x 256 pixels x 2 passes (NEX)
GRE/30 Multi (Grass Echo)
28cm field of view
3mm thickness
0mm interscan skip
8 images

Vowel	TR	TE	Time
i	32	13	2:25
ɪ	37	11	2:36
e	39	13	2:44
ɛ	39	12	2:44
u	37	11	2:36
ʊ	37	11	2:36

Subject MT (English)

256 x 256 pixels x 2 passes (NEX)
GRE/30 Multi (Grass Echo)
28cm field of view
3mm thickness
0mm interscan skip
8 images

Vowel	TR	TE	Time
i	36	10	2:32
ɪ	36	10	2:32
e	37	11	2:32
ɛ	37	11	2:32
u	36	10	2:32
ʊ	37	11	2:32

Table 2. *Axial imaging parameters*

Subject AO (Akan)
256 x 128 pixels x 1 pass (NEX)
SPGR/45 Volume (Spoiled grass)
28cm field of view
5mm thickness
0mm interscan skip
28 images
TR 45
TE 5
Time 3:05
Subject MT (English)
256 x 128 pixels x 1 pass (NEX)
SPGR/45 Volume (Spoiled grass)
30cm field of view (i, I, u, U)
28cm field of view (e, E)
5mm thickness
0mm interscan skip
28 images
TR 45
TE 5
Time 3:05

Analysis

The images obtained were converted from 16 bit Signa format to an 8 bit (255 gray level) format compatible with display and analysis software. The axial images were normalized for a standard image density so that a given pixel magnitude represented the same value across all series; this was done so that air-tissue boundary measurements could be made consistently. Image analysis was performed on a Macintosh II computer using the NIH IMAGE program.[6]

The sagittal images were used to replicate the cineradiographic results. Measurements of relative height for the tongue dorsum, jaw, and larynx were obtained for each vowel, using the top of the second cervical vertebra as the reference, from the image judged to be closest to sagittal midline for that scan set. Dorsum height was measured from the point on the tongue that maximized the vertical distance from the reference (see Figure 1); jaw height was measured from a point at the base of the root of the lower incisors; and larynx height was measured from its vertical distance to the reference. Relative tongue root advancement was measured as the length of a line drawn from the bottom of the vallecular sinuses to the rear pharyngeal wall. Figure 1 illustrates how measurements were obtained, and shows an overlay of the vocal tract outlines obtained for the comparison between Akan /i/ and /ɪ/.

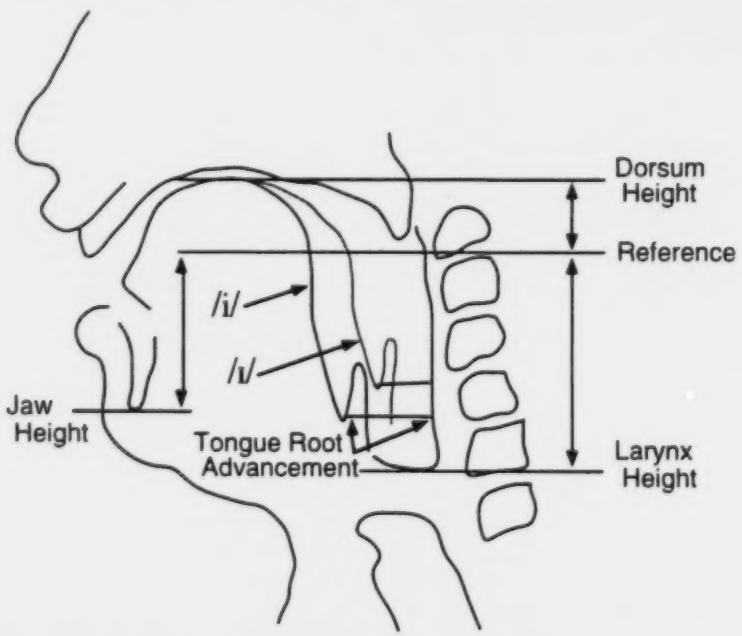


Figure 1. **Sagittal measurements, showing Subject AO (Akan) for /i/ overlain by /ɪ/.**

The axial images were used to obtain data for the left-to-right lateral dimension inaccessible to the cineradiographic studies. For purposes of comparison 11 sequential images were chosen from each axial scan set. The reference used for comparison was the lowest image in each series in which the epiglottis was visible as a detached entity; this point corresponds to the bottom of the vallecular sinuses used in the sagittal images to identify the tongue root position. This image was used as the pivot for choosing five sequential images below that level in the pharynx, and five more above, for the total of eleven. The approximate locations of axial scans measured are shown overlain on a sagittal outline in Figure 2. The reason for using the epiglottis to coordinate inter-vowel comparison was to minimize any effects due to laryngeal height differences, so that comparison would be made at corresponding anatomical points for both vowel conditions.

Three measurements were obtained from each axial image: the distance from the anterior to the posterior boundaries of pharyngeal airspace corresponding to tongue root advancement, which will be referred to as "depth"; the novel left-to-right lateral distance across pharyngeal airspace, which will be referred to as "width"; and a measure of the cross-sectional area of pharyngeal airspace at that level; these are illustrated in Figure 3. The distance measurements were made along a straight line at the point of widest distension for each dimension. The area measurement was computed by determining a perimeter of standard image intensity corresponding to the pharyngeal air-tissue boundary, counting the number of enclosed pixels, and converting to area measure. Except when determining the reference or pivot level, the epiglottis was ignored in all measurements; in particular at levels above the reference (those in which a detached epiglottis was visible), the anterior wall was taken to be the base of the tongue, and area measurements included any visible epiglottal tissue.

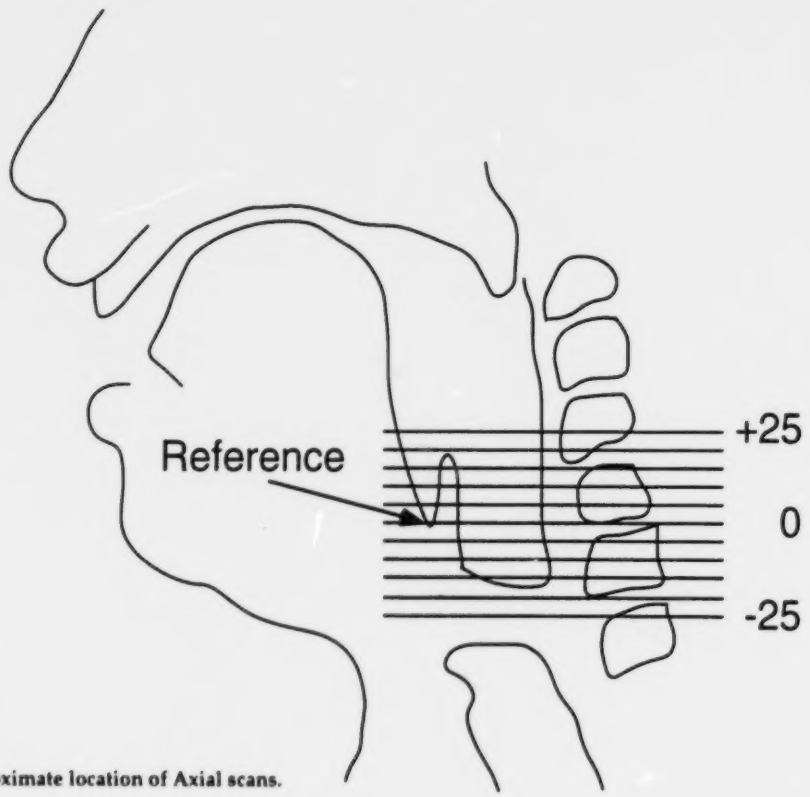


Figure 2. **Approximate location of Axial scans.**

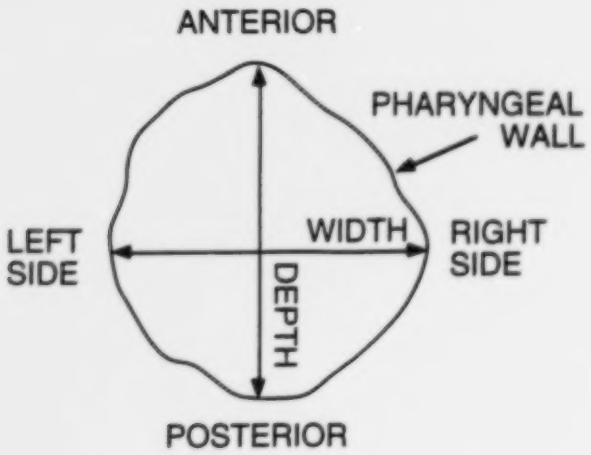


Figure 3. **Axial measurements.**

Axial image measurement is complicated by the trifurcation of the airspace below the epiglottis into three tubes by the aryepiglottic folds: the central laryngeal passageway and two deadend sidechannels, the piriform sinuses. The approach taken for these images was to obtain an overall area measurement by adding the cross-section measured for the central larynx tube to those obtained for each of the piriform sinuses (i.e. aryepiglottic tissue was not included). The width measurement was made from the left side of the left sinus to the right side of the right sinus, but the depth measurement was made across the point of widest distension of the central larynx tube only.

The acoustic recordings of pre- and post-scan phonation were used to verify that subject vowel quality remained on target during scanning. The resulting utterances were digitized at a sampling frequency of 10 kHz. Formant values for each token as well as the original vowel target utterances were obtained by a linear-predictive-coding autocorrelation analysis, using a 20ms Hamming window shifted 10ms to fit a 14th-order polynomial, with averaging of successive frames for which the first three formants were available.

Results

Images. Figure 4 is representative of the resulting images. The left side of the figure shows a midline sagittal view and mid-pharynx view from the expanded (+ATR) variant of the Akan vowel /i/, and on the right the corresponding constricted views (/ɪ/). The orientation of the axial images corresponds to a slice through the neck, perpendicular to the pharynx, presented so that the top of the image shows the front of the face or neck. Notice the difference in pharyngeal airspace: the sagittal view on the left shows a wider opening in the lower pharynx than does the corresponding view for the constricted variant on the right. Similarly the axial view of the expanded vowel on the left shows a larger cross-sectional area than the corresponding constricted vowel on the right.

In subjective terms the images obtained varied from poor to good quality compared to others of this type,[7] representing in general a reasonable compromise between image noise and required scanning time. Measurements were based on air/tissue interfaces, which tended to image well even if resolution of tissue-internal anatomical details was nonoptimal. Resolution was approximately one millimeter per image pixel in both orientations.[8]

To determine whether normalization of measurements for different head sizes was appropriate, subject tract lengths were measured using the midline sagittal image obtained for the vowels /i/ and /ɪ/ in each language. The measurement was made from the level of the vocal folds, around the curve of the tongue, to the midpoint of a line bisecting the narrowest lip approximation, presented here in centimeters:

(5) Vocal Tract Lengths (cm)

Vowel	Subject AO (Akan)	Subject MT (English)
i	17.4	16.9
ɪ	15.7	15.3

These values were considered to be sufficiently close as to make normalization unnecessary.

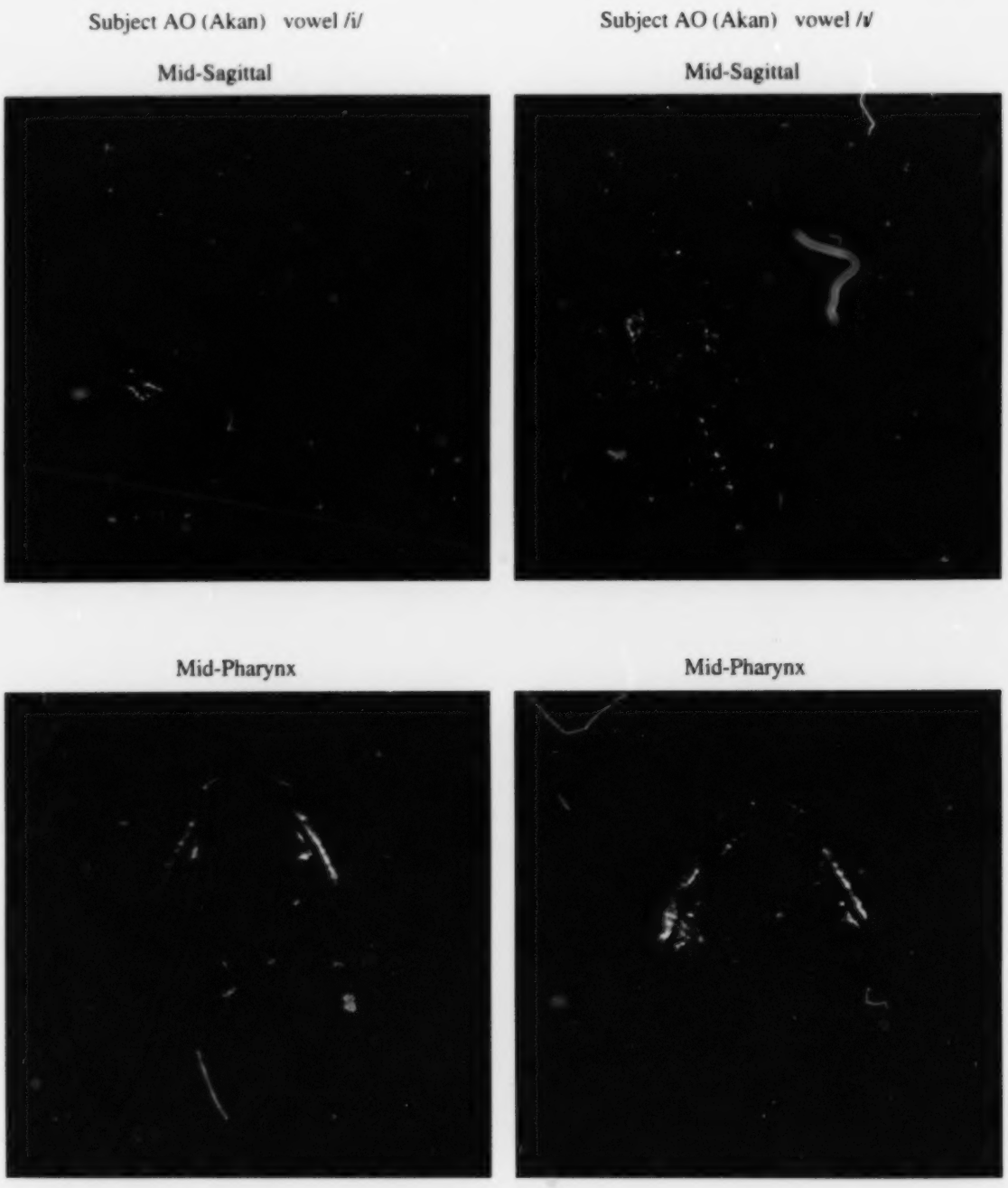


Figure 4. **Representative images.**

Errors affecting measurements may have been introduced in two different ways: from distortion of the image due to subject movement during scanning, or from subject head tilt. Both possibilities affected sagittal scanning more than axial. Sagittal scans were assembled individually one after another, whereas the axial scans were obtained using a method that computed all images

BEST COPY AVAILABLE

concurrently. Subject movement in a sagittal scan set resulted in blurring of the single image being acquired at that moment, making it unusable for measurement, while movement in an axial set caused only a loss of definition across all images in the set, leaving them all still viable for measurement.

The effect of any head tilt or rotation was to cause the scanning plane to intersect with the subject at an oblique angle, rather than producing a true bisection of the head. This was not a problem for axial images, since any tilting would have been too slight to cause measurable distortion in that orientation. For the sagittal images however, even slight tilting was sufficient to make determination of the head midline difficult; for example, one image might show midline at the level of the larynx, yet be off center at the level of the palate, thus affecting tongue dorsum height measurements. Therefore all sagittal tongue measurements were confirmed on images adjacent to the one chosen as midline.

Sagittal Orientation. The sagittal measurements obtained are given in Table 3, and illustrated graphically in Figure 5. Figure 6 shows the (expanded - constricted) difference between results obtained for corresponding vowel pairs.

Table 3. *Sagittal measurements.*

Subject AO (Akan)		(mm)		
Vowel	Root Advancement	Dorsum Height	Laryngeal Depth	Jaw Depth
i	22.97	19.69	68.91	48.13
ɪ	17.50	19.69	57.97	45.94
e	21.88	25.16	54.69	41.56
ɛ	19.69	24.06	51.41	47.03
u	32.81	21.88	74.38	43.75
ʊ	17.50	18.59	67.81	41.56

Subject MT (English)		(mm)		
Vowel	Root Advancement	Dorsum Height	Laryngeal Depth	Jaw Depth
i	22.97	30.63	55.78	30.63
ɪ	21.88	28.44	47.03	32.81
e	21.88	27.34	51.41	36.09
ɛ	19.69	25.16	49.22	32.81
u	30.63	30.63	55.78	33.91
ʊ	18.59	26.25	54.69	33.91

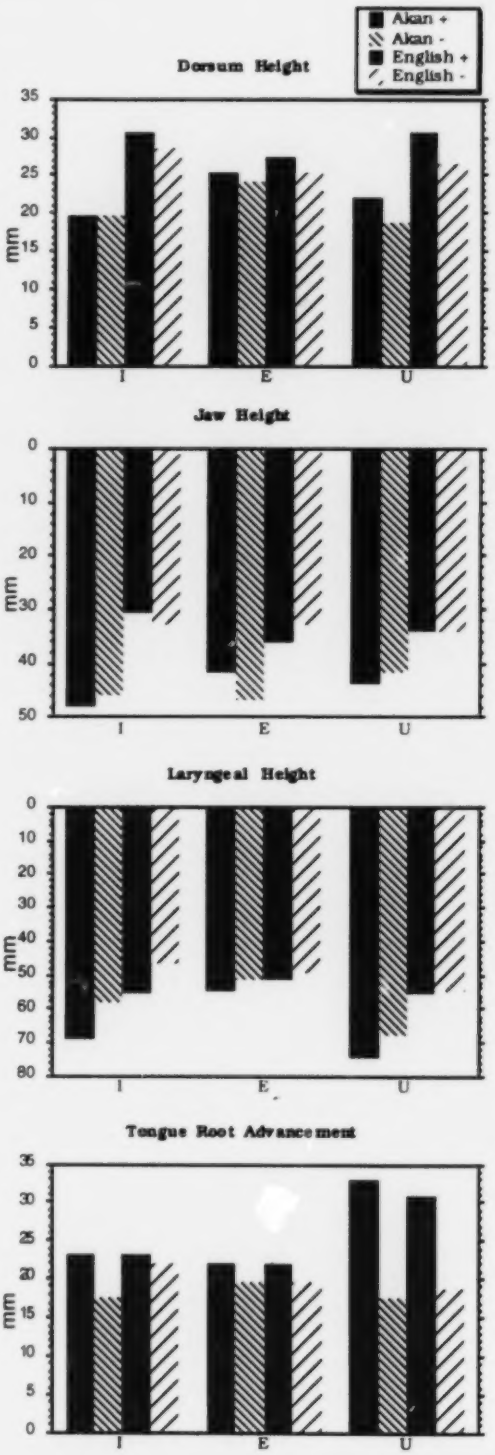


Figure 5. **Sagittal results.**

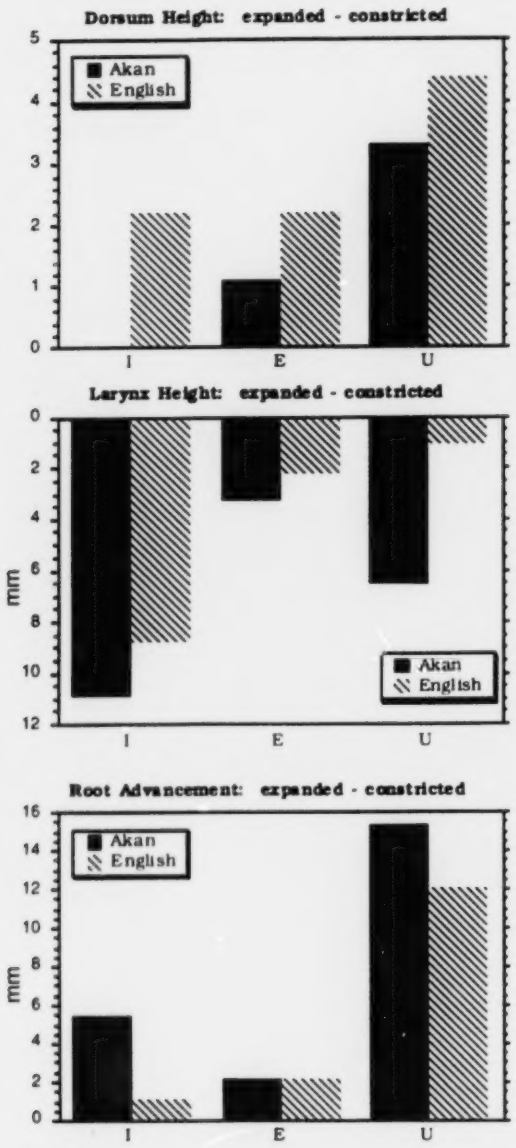


Figure 6. Expanded - constricted sagittal results.

In each case the Akan results confirm the cineradiographic studies mentioned above: tongue root advancement was larger for the expanded (+ATR) variant of each vowel pair, larynx height was lower for each expanded variant as well, and tongue dorsum height showed either no difference (i:ɪ) or slightly higher values for expanded alternates (e:ɛ, u:ʊ). The results for English were very similar: for each tense variant root advancement was larger, larynx height was lower, and tongue dorsum higher. However, the English measurements show smaller differences in magnitude for both root advancement and laryngeal lowering than Akan, and greater differences in tongue dorsum height. Results obtained for jaw height showed small and inconsistent differences in both languages.

Axial Orientation. The axial measurements obtained are given in Tables 4a and 4b. Figure 7 illustrates the difference in cross-sectional area between Akan vowel pairs. In each graph in this and subsequent figures the height of each bar shows the difference in area between expanded and constricted variants at corresponding levels of the pharynx, increasing in height at 5mm intervals from five measurement levels below the base of the epiglottis to five levels above. Observe that the pharyngeal airspace is larger at all measured levels for the expanded (+ATR) variants of vowels /i/ and /u/, and for all levels of expanded /e/ above the epiglottal pivot. By combining each sequence of measured cross-sections the following approximations to pharyngeal volume were obtained (representing the section delimited by +/- 25mm from the base of the epiglottis):[9]

(6) Derived Pharyngeal Volumes for Subject AO (Akan) (cm^3)

	I	E	U
+ATR	30.58	18.38	42.04
-ATR	18.90	17.27	25.09

The larger volumes observed for the "Expanded" variants of each vowel show that Lindau's term is an apt name for the feature being contrasted. It is interesting that the ratio of expanded to constricted volume is nearly the same for vowels /i/ and /u/ (1.62 *vs.* 1.68), and the absence of a difference of similar magnitude for /e/ suggests that the data obtained for that vowel should be viewed with caution (although they agree with the sagittal results, obtained during a separate scanning sequence, which also showed the smallest root advancement difference for /e/). The smaller differences observed for /e:ɛ/ may be due to the fact that the inherently lower tongue constriction location characteristic of these vowels leaves less room in the lower pharynx for effecting the contrast.

Table 4a. *Axial measurements for Subject AO (Akan).*

Area (mm^2)

Level (mm)	i	ɪ	e	ɛ	u	ʊ
-25	309.84	133.98	69.38	107.67	508.42	245.24
-20	230.88	137.57	137.57	205.76	534.74	290.70
-15	296.68	230.88	174.66	293.09	807.50	363.67
-10	419.90	373.24	294.29	327.78	820.65	504.83
-5	492.87	313.43	328.98	311.04	788.35	520.39
0	656.76	422.29	476.12	297.88	1233.37	878.08
5	750.07	451.00	488.09	374.44	1165.19	764.43
10	787.16	455.79	462.96	380.42	721.36	480.91
15	722.56	437.84	429.47	418.70	646.00	381.62
20	714.18	421.09	404.35	374.44	640.01	349.32
25	734.52	403.15	410.33	362.48	543.12	238.06

Width (mm)

Level (mm)	i	ɪ	e	ɛ	u	ʊ
-25	35.00	28.44	28.44	28.44	33.91	29.53
-20	32.81	29.53	29.53	29.53	35.00	30.63
-15	32.81	32.81	31.72	30.63	39.38	31.72
-10	35.00	33.91	33.91	29.53	38.28	31.72
-5	36.09	31.72	35.00	30.63	37.19	36.09
0	32.81	29.53	33.91	25.16	37.19	38.28
5	30.63	24.06	29.53	24.06	35.00	32.81
10	30.63	26.25	22.97	25.16	35.00	30.63
15	28.44	28.44	28.44	24.06	35.00	27.34
20	31.72	25.16	24.06	22.97	35.00	27.34
25	30.63	22.97	25.16	17.50	33.91	20.78

Depth (mm)

Level (mm)	i	ɪ	e	ɛ	u	ʊ
-25	19.69	18.59	20.78	12.03	37.19	26.25
-20	25.16	15.31	20.78	14.22	31.72	21.88
-15	27.34	18.59	19.69	15.31	31.72	24.06
-10	25.16	17.50	24.06	15.31	29.53	22.97
-5	17.50	14.22	14.22	14.22	30.63	20.78
0	33.91	24.06	27.34	17.50	43.75	30.63
5	33.91	18.59	27.34	20.78	42.66	28.44
10	31.72	18.59	18.59	22.97	32.81	21.88
15	29.53	24.06	20.78	21.88	27.34	18.59
20	32.81	22.97	24.06	24.06	27.34	15.31
25	31.72	24.06	24.06	25.16	25.16	12.03

Table 4b. *Axial measurements for Subject MT (English).*

Area (mm^2)

Level (mm)	i	ɪ	e	ɛ	u	ʊ
-25	438.08	508.12	470.14	486.89	753.94	618.48
-20	517.73	519.10	411.52	514.40	699.01	617.29
-15	527.34	616.61	391.19	614.89	806.12	608.91
-10	661.93	649.57	576.61	590.97	862.43	752.47
-5	606.99	585.02	523.97	644.80	896.76	718.97
0	914.61	659.18	608.91	532.35	1031.34	848.17
5	951.69	630.34	653.17	368.46	958.56	802.71
10	942.08	611.11	575.42	295.48	944.82	724.95
15	948.94	597.38	517.99	324.19	855.56	628.05
20	940.70	649.57	526.37	337.35	811.61	541.92
25	961.30	708.62	525.17	394.78	752.56	485.69

Width (mm)

Level (mm)	i	ɪ	e	ɛ	u	ʊ
-25	32.81	33.98	33.91	33.91	33.98	32.81
-20	37.50	36.33	36.09	36.09	36.33	33.91
-15	39.84	39.84	38.28	38.28	39.84	36.09
-10	41.02	39.84	39.38	41.56	41.02	39.38
-5	41.02	41.02	39.38	40.47	41.02	40.47
0	38.67	38.67	40.47	38.28	39.84	39.38
5	37.50	38.67	39.38	37.19	38.67	39.38
10	42.19	36.33	36.09	36.09	39.84	39.38
15	43.36	41.02	35.00	33.91	41.02	31.72
20	44.53	38.67	36.09	35.00	41.02	36.09
25	45.70	42.19	33.91	33.91	39.84	38.28

Depth (mm)

Level	i	ɪ	e	ɛ	u	ʊ
-25	26.95	29.30	29.53	26.25	30.47	30.63
-20	21.09	22.27	21.88	22.97	28.13	28.44
-15	19.92	19.92	21.88	24.06	29.30	26.25
-10	21.09	19.92	17.50	19.69	24.61	29.53
-5	17.58	17.58	17.50	17.50	24.61	24.06
0	31.64	23.44	24.06	17.50	31.64	26.25
5	30.47	22.27	22.97	15.31	30.47	27.34
10	30.47	21.09	20.78	14.22	29.30	19.69
15	30.47	18.75	20.78	12.03	25.78	18.59
20	31.64	23.44	18.59	12.03	24.61	16.41
25	29.30	23.44	20.78	17.50	21.09	15.31

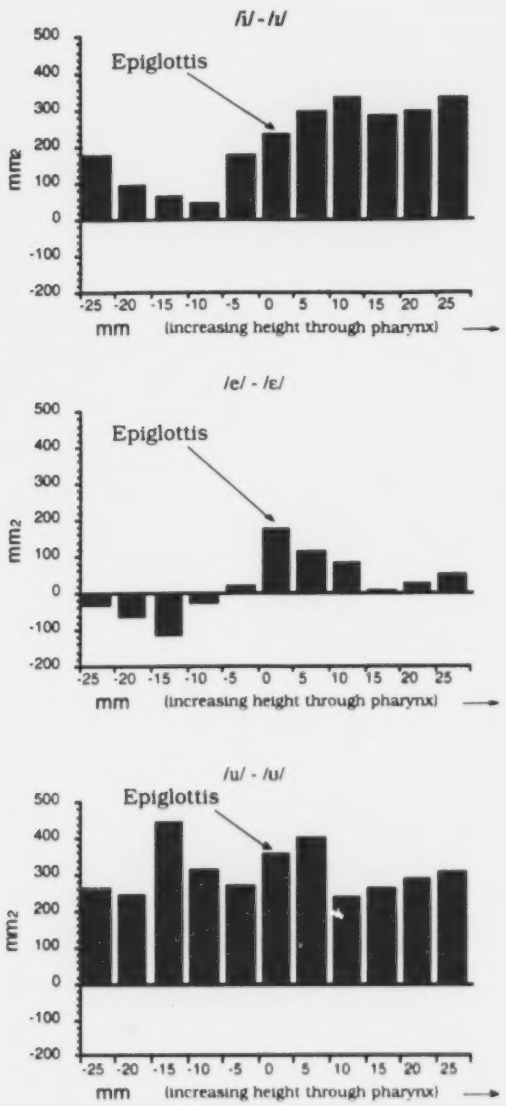


Figure 7. **Delta cross-sectional area for Subject AO (Akan).**

The observed difference in pharyngeal volume is consistent with the tongue root advancement observed in the sagittal images and cineradiographic studies, but leaves open the question of whether the difference is due entirely to root position, or whether pharyngeal lateral width contributes as well. The relative contributions of width and depth to the area values obtained for Akan are illustrated in Figure 8.

Recall that width is a measure of side-to-side or lateral distance, and depth is a measure of anterior/posterior distance corresponding to root advancement. Notice that while somewhat noisier than the area data, the overall trend for both width and depth is for consistently larger values for the expanded variant of each vowel.[10] Notice also that the observed differences in lateral width are almost as large as those observed for anterior/posterior depth, showing that for this speaker at least control of the lateral dimension is an important part of the mechanism used to produce the contrast, thereby confirming Lindau's position that speakers of Akan seek to maximize the difference in overall pharyngeal volume in effecting the contrast.

Figures 9, 10, and 11 provide a comparison of the English area, width, and depth results with those of Akan. As in the previous figures, the height of each bar shows the difference between expanded and constricted variants at corresponding levels through the pharynx. Combining area cross-sections as for Akan the following approximations to pharyngeal volume were obtained:

(7) Derived Pharyngeal Volumes for Subject MT (English) (cm^3)

	I	E	U
Tense	37.25	26.28	43.10
Lax	30.13	23.55	34.31

As in Akan, the expanded (tense) variants show consistently larger overall pharyngeal volume than their lax counterparts. Again, it is interesting to note that the ratio of expanded to constricted volume is nearly the same for vowels /i/ and /u/ (1.24 *vs.* 1.26), although the magnitude of the difference is considerably less than that observed for Akan.

Acoustics. Figure 12 shows a combined vowel space chart using the averaged stimulus target vowel formants from each language. The primary acoustic effect of the ATR contrast on Akan vowel pairs appears to be a raised F1 for [-ATR] variants, with F2 relatively unaffected. Lax vowels in English on the other hand show both a raised F1 and lowered F2 compared to their tense counterparts, for a net centralizing effect. Values measured for F3 were lower for constricted variants in both languages. Two analyses of variance were performed to quantify the significance of these observed effects: one grouped recorded tokens into source groupings of target (the original stimulus words), sagittal run, and axial run; the second analysis grouped them into target, pre-scan, and post-scan source groupings.

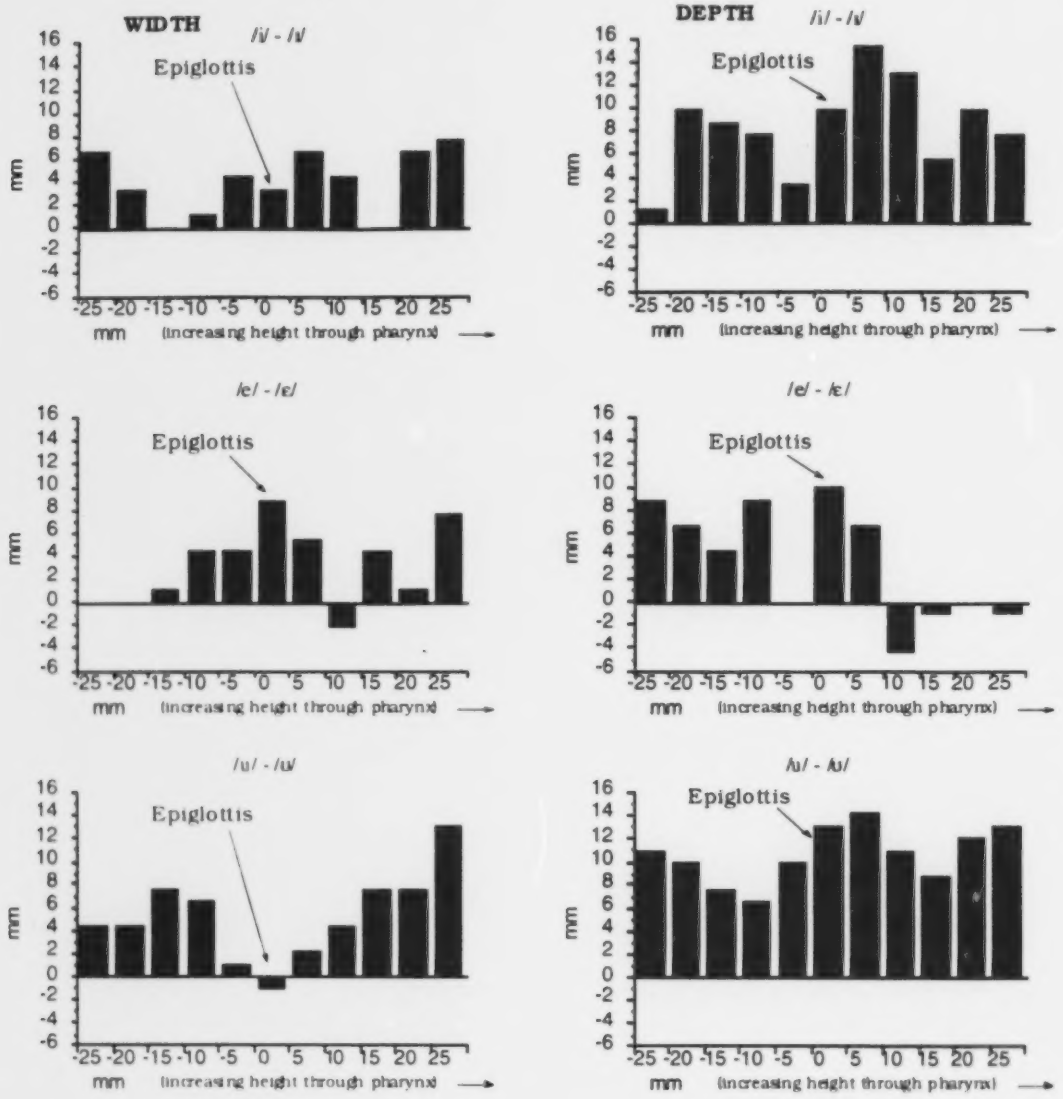


Figure 8. **Akan pharyngeal width and depth.**

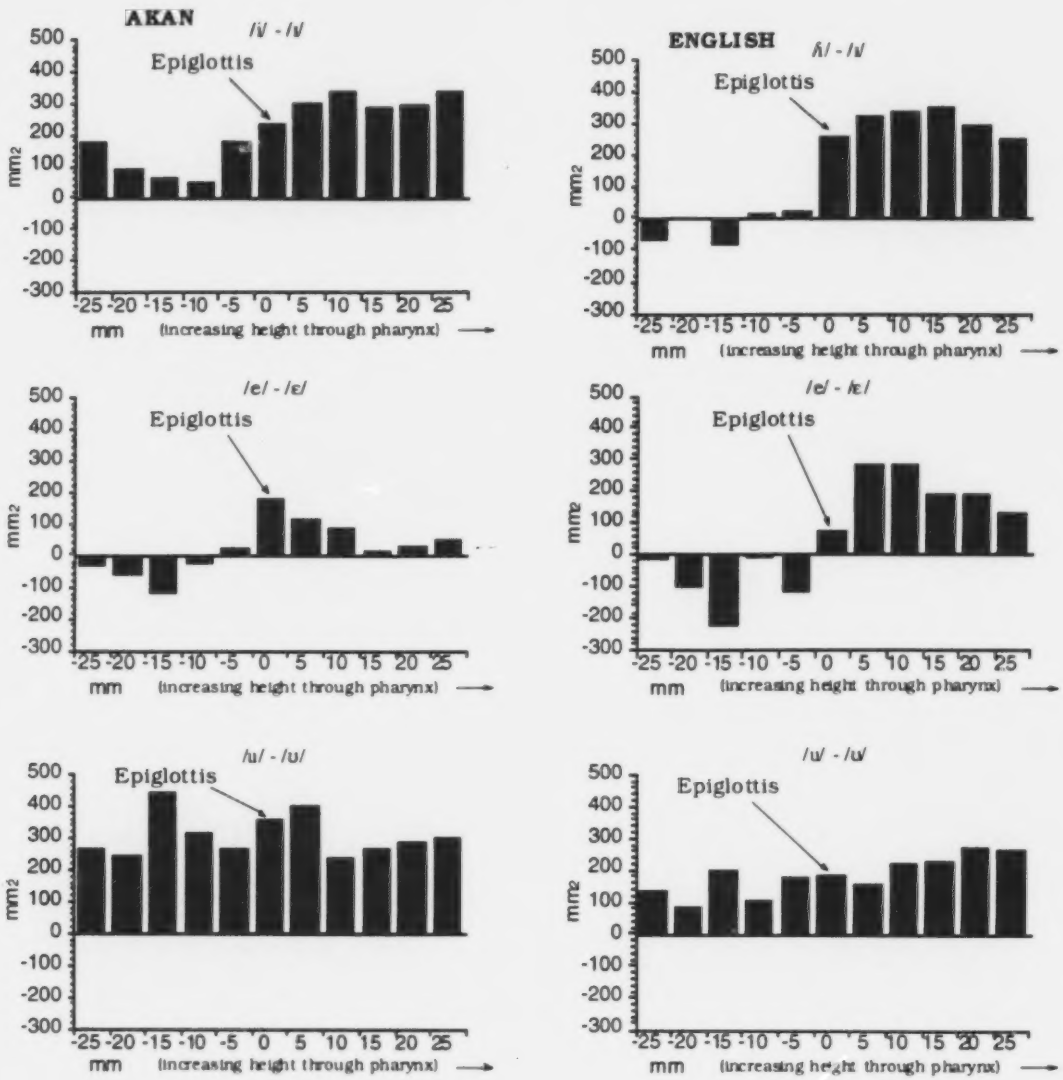


Figure 9. **Comparison of cross-sectional area.**

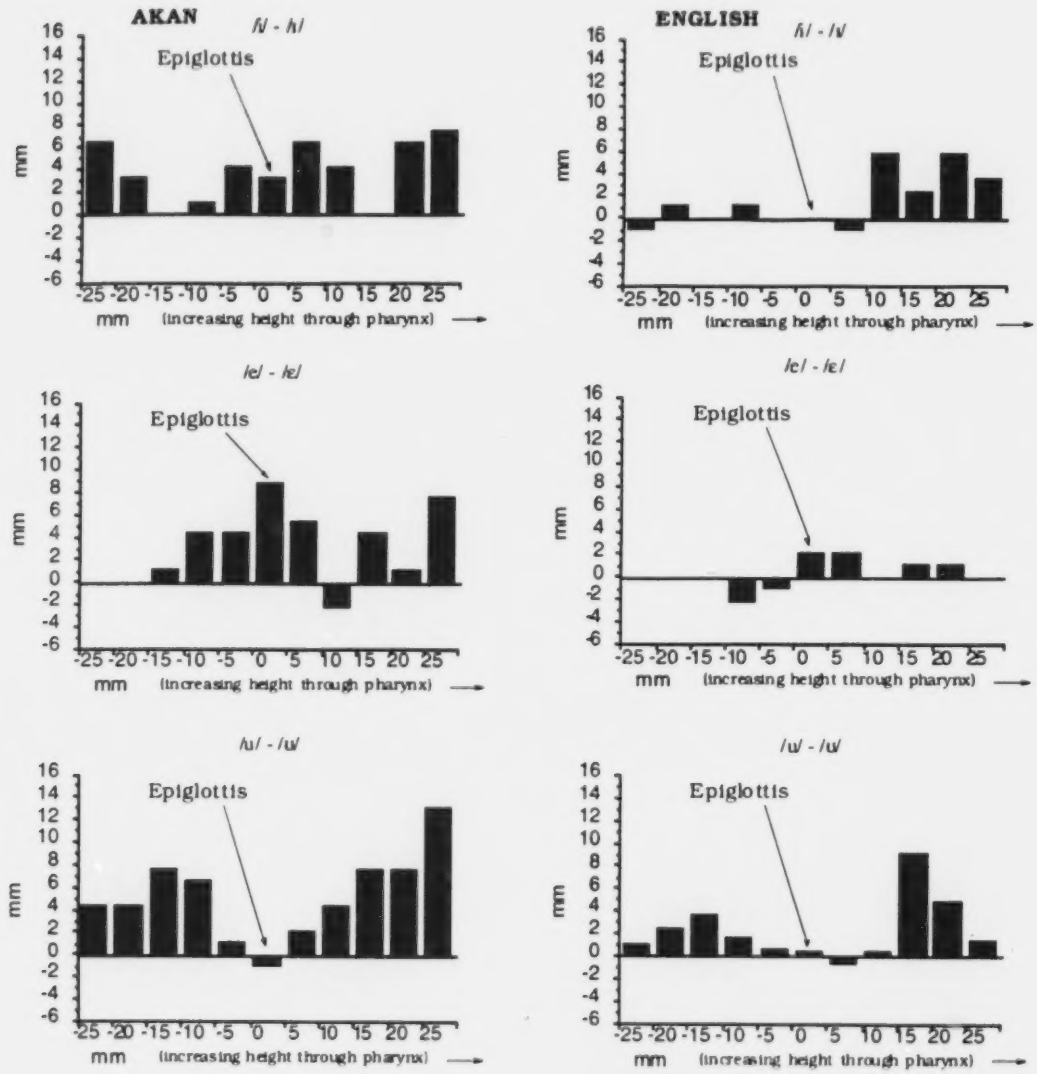


Figure 10. **Comparison of pharyngeal width.**

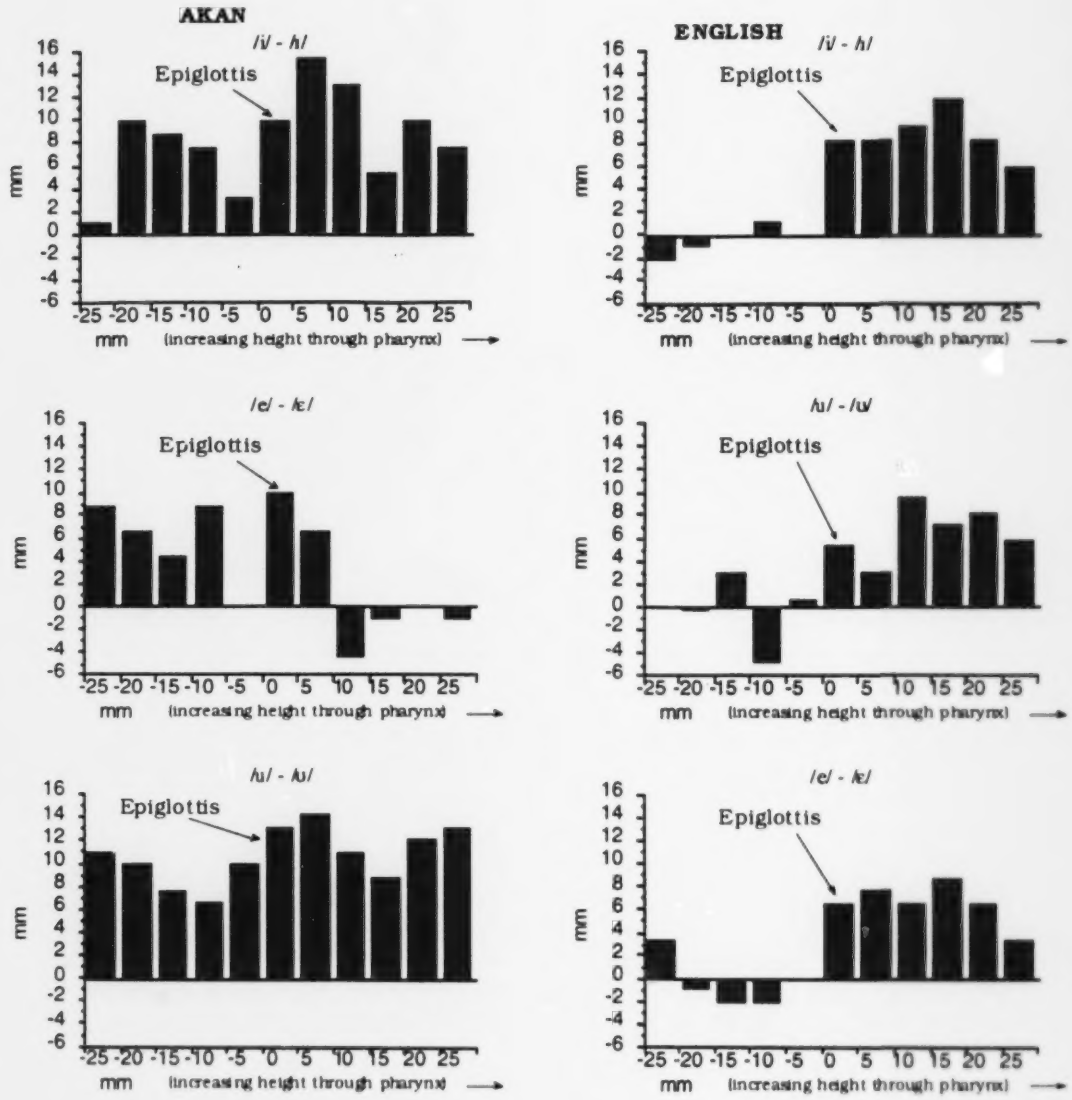


Figure 11. **Comparison of pharyngeal depth.**

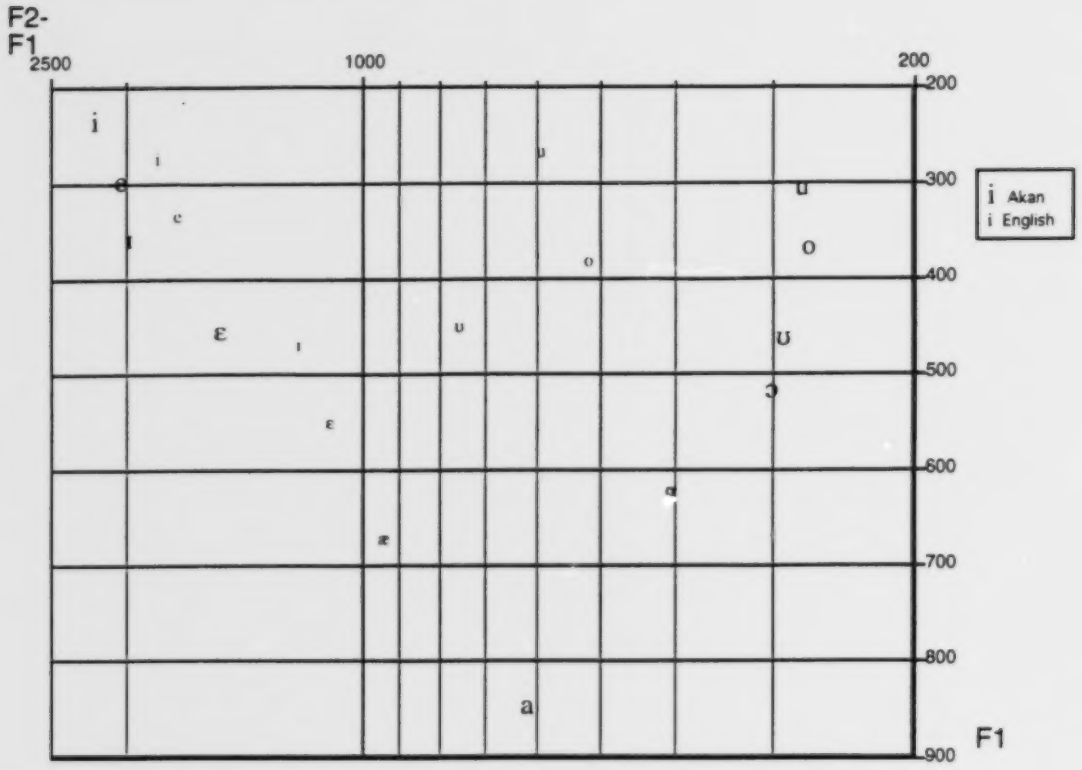


Figure 12. **Combined vowel space.**

The results of both analyses confirmed the significance of the ATR and Tense/Lax effects on F1 ($p < .0001$) and F3 ($p < .05$), showed strong significance for the effect of the Tense/Lax contrast on F2 ($p < .0001$), and showed no significance for ATR effects on F2.

Figures 13 and 14 show vowel space plots for subjects AO (Akan) and MT (English) respectively incorporating both stimulus and run time formant values. Missing values for certain configurations resulted from subjects initiating phonation after scanning had commenced, or respiring as it ended (recall that machine noise precluded analysis of phonation recorded during scanning). Ideally the formant values measured for a given vowel should be independent of its source category; in practice this expectation was unrealistic, given that the target vowels were separately recorded under acoustically ideal conditions, while pre- and post-scan tokens were recorded with subjects prone and encased by several tons of intimidating high technology.[11] Effects due to token source on measured formant values were assessed by computing Wilks' Λ (a measure of multivariate correlation) from the separate formant analyses. The values obtained (listed in Table 5) show that token source category was in fact a significant source of variance for both subjects, both as a main effect, and in interactions with vowel and ATR factors, but the results were mixed depending on how the runtime tokens were grouped. When analyzed by type (i.e. target, pre-scan, post-scan), the Akan subject showed a marginally significant ($p < .05$) interaction of source with ATR (the feature of primary interest), but no significance for this interaction when analyzed by run (i.e. target, sagittal, axial). This suggests that while subject AO drifted slightly from the initial target over time, the way in which he did so was consistent across runs, at least with respect to ATR effects. The English subject MT showed the reverse pattern: the interaction of source with the Tense/Lax parameter was marginally significant ($p < .05$) for the analysis of tokens grouped by run, but not significant grouped by type, indicating that while this subject produced slightly different vowels across runs, he did not drift significantly during scanning.

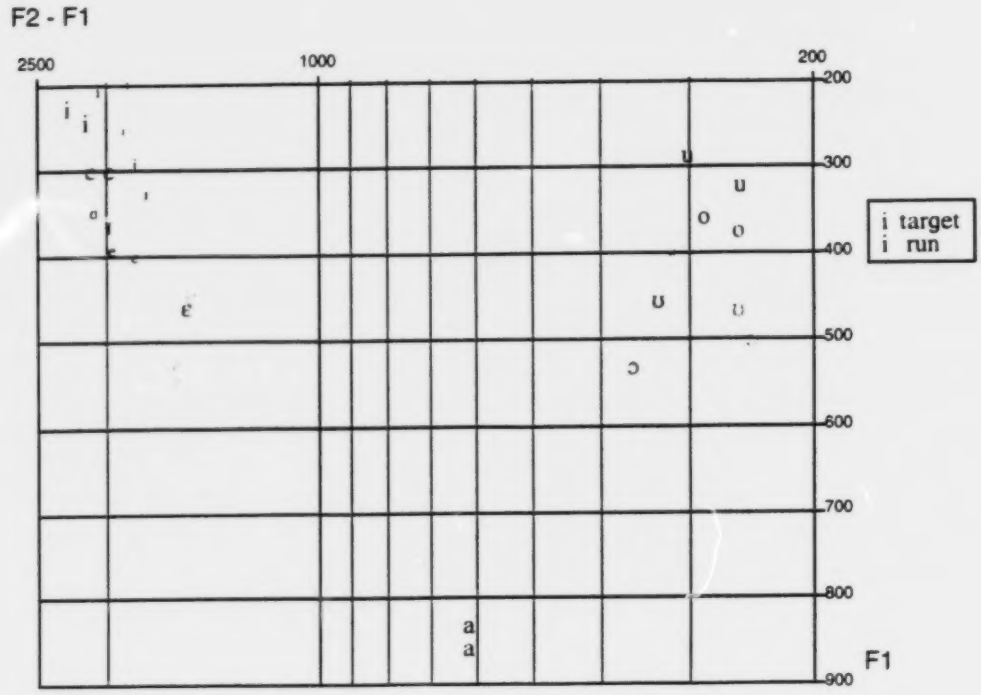


Figure 13. **Subject AO (Akan) vowel space.**

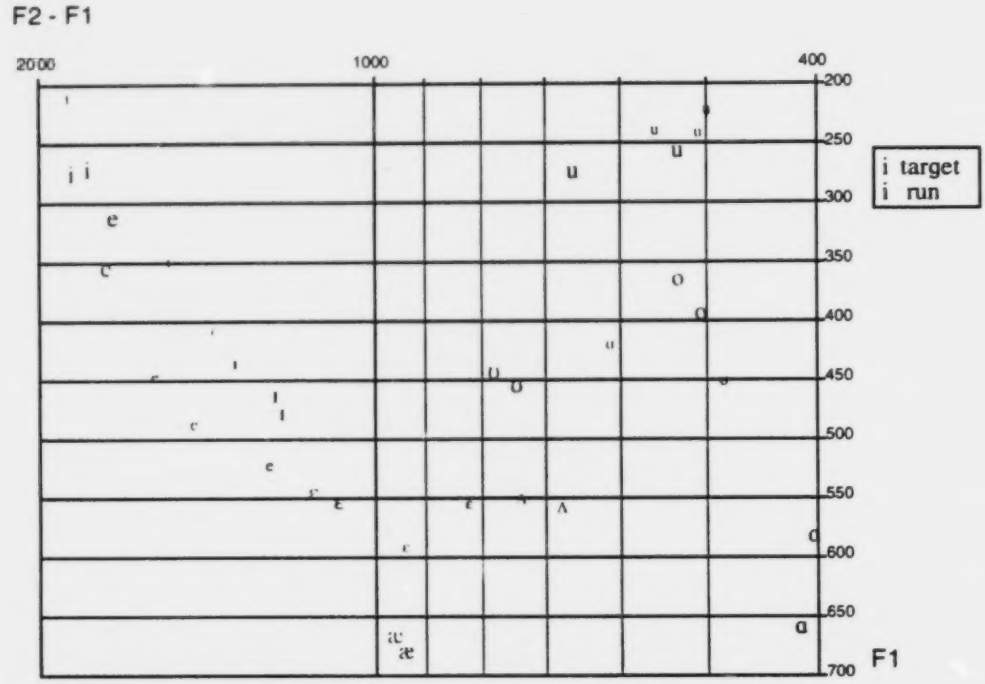


Figure 14. **Subject MT (English) vowel space.**

Table 5. *Wilks' Lambda values from MANOVA of formant data. Tokens grouped as Target:Sagittal:Axial*

Akan

Effect	Value	F-Value	Num DF	Den DF	P-Value
Vowel	0.001	94.142	6.000	18.000	0.0001
ATR	0.112	23.867	3.000	9.000	0.0001
Source	0.223	3.348	6.000	18.000	0.0215
Vowel * ATR	0.063	8.951	6.000	18.000	0.0001
Vowel * Source	0.153	2.847	9.000	22.054	0.0206
ATR * Source	0.414	1.665	6.000	18.000	0.1871
Vowel * ATR * Source	0.415	1.656	6.000	18.000	0.1893

English

Effect	Value	F-Value	Num DF	Den DF	P-Value
Vowel	0.001	136.776	6.000	28.000	0.0001
ATR	0.055	80.075	3.000	14.000	0.0001
Source	0.410	2.619	6.000	28.000	0.0383
Vowel * ATR	0.067	13.420	6.000	28.000	0.0001
Vowel * Source	0.149	3.281	12.000	37.332	0.0025
ATR * Source	0.369	3.020	6.000	28.000	0.0210
Vowel * ATR * Source	0.456	1.074	12.000	37.332	0.4078

Tokens grouped as Target:Pre:Post scan

Akan

Effect	Value	F-Value	Num DF	Den DF	P-Value
Vowel	0.001	140.397	6.000	20.000	0.0001
ATR	0.054	58.871	3.000	10.000	0.0001
Source	0.086	8.057	6.000	20.000	0.0002
Vowel * ATR	0.080	8.483	6.000	20.000	0.0001
Vowel * Source	0.172	2.889	9.000	24.488	0.0174
ATR * Source	0.268	3.106	6.000	20.000	0.0257
Vowel * ATR * Source	0.684	1.542	3.000	10.000	0.2639

English

Effect	Value	F-Value	Num DF	Den DF	P-Value
Vowel	0.000	205.738	6.000	28.000	0.0001
ATR	0.060	72.659	3.000	14.000	0.0001
Source	0.167	6.748	6.000	28.000	0.0002
Vowel * ATR	0.016	32.756	6.000	28.000	0.0001
Vowel * Source	0.058	6.016	12.000	37.332	0.0001
ATR * Source	0.520	1.805	6.000	28.000	0.1343
Vowel * ATR * Source	0.304	1.765	12.000	37.332	0.0907

Recall that the purpose of collecting acoustic data was to verify subjects were in fact articulating the desired vowel configuration throughout the scanning sequence. The above results indicate that due possibly to nervousness or fatigue subjects were not in fact entirely consistent in vowel production. However, if the F-values derived using Wilks' Λ are regarded as reflecting the relative importance of the individual factors, observe that vowel type and ATR and their interaction show larger values than any term involving token source category, for both languages and both analyses; in other words while the source factor is significant, it is less important relative to the other parameters determining vowel configuration, and does not indicate subject phonation drifted so much as to seriously skew the imaged tract configurations.

Discussion

The fact that the English sagittal measurements show smaller differences in magnitude for both tongue root advancement and laryngeal lowering than Akan, and greater differences in tongue dorsum height, suggests a relatively more significant role for tongue height in maintaining the English contrast. It should be noted however that root advancement and dorsum height may be intrinsically linked. An electromyographic study by Baer, Alfonso, and Honda (1988) claimed that

the same contraction of the posterior genioglossus (GGP) muscle effecting tongue root advancement also forces the tongue dorsum upwards, while contraction of the separately controlled anterior genioglossus (GGA) pulls the dorsum forward and down. The results obtained here suggest that GGA activity is greater in Akan than in English: active control of the GGA in Akan works against the dorsum-raising effect of the GGP, resulting in smaller net dorsum height differences than those observed for English. This predicts that the correlation of dorsum height with root advancement should be stronger in English than in Akan, which is confirmed by a regression analysis outlined in the table below:

(8) Dorsum Height regressed against Root Advancement (all vowels)

	Subject AO (Akan)	Subject MT (English)
d.f.	4	4
Pearson's r	0.226	0.782
r^2	0.051	0.611
t-ratio	0.465	2.51
p	n.s.	<0.05

Assuming the pattern of muscle activity mentioned above, one apparent difference therefore between the ATR and Tense/Lax mechanisms lies in how each language treats the inherent dorsum-raising effect of root advancement: Akan seeks to neutralize its effect, whereas English (at least in this subject's dialect) appears to exploit it.

Another difference is apparent from the patterning of axial data at measured levels below the epiglottis. With one exception (area measured at the three lowest levels of /e/), the area, width, and depth measurements obtained for Akan show consistently larger values for expanded (+ATR) variants at all measured levels, above and below the epiglottis. But while the English data also show consistently larger values above the epiglottal pivot, at levels below that point differences between tense and lax variants are inconsistent in sign and considerably smaller in magnitude. The change occurs abruptly, and is evident to some degree in all three parameters measured.

An attempt was made to quantify the significance of this divergence by performing an analysis of variance on each measurement parameter, treating the individual levels as repeated measures nested within an upper or lower grouping factor.[12] Under this design, inter-language differences in behavior above and below the epiglottis between expanded and constricted configurations are encompassed in the Language × Group × ATR interaction. Results of the analysis for the width parameter showed no significance for this interaction (F=1.12), but moderate significance for depth (F=9.48, $p<.05$), and strong significance for area (F=15.84, $p<.005$), reflecting the difference in subepiglottal behavior between the two subjects.

Further evidence of divergent behavior is provided by a regression analysis of width against depth for corresponding measurement levels, summarized in Table 6, and illustrated in Figure 15. The results show a significant ($p < .01$) positive correlation between Akan width and depth across all measurement levels, but no correlation for the same analysis on English. When the English values are analyzed separately by upper/lower grouping however, the results show a significant ($p < .01$) positive correlation between width and depth for measurement levels above the epiglottal pivot, and a significant ($p < .01$) *negative* correlation at levels below it; in other words below the level of the epiglottis in English an increase in anterior/posterior depth is accompanied by a decrease in lateral width. Separate analysis of the Akan values by group shows significant ($p < .01$) positive correlations between width and depth for measurement levels above and below the epiglottal pivot.

Table 6. *Regression of Axial Width and Depth.*

All Vowels	Subject AO (Akan)	Subject MT (English)
d.f.	58	58
Pearson's r	0.443	0.090
r^2	0.196	0.008
t-ratio	3.76	-0.687
p	<0.01	n.s.
Akan Vowels (Subject AO)	**below epiglottis**	**above epiglottis**
d.f.	28	28
Pearson's r	0.545	0.744
r^2	0.297	0.553
t-ratio	3.44	5.89
p	<0.01	<0.01
English Vowels (Subject MT)	**below epiglottis**	**above epiglottis**
d.f.	28	28
Pearson's r	-0.671	0.680
r^2	0.450	0.462
t-ratio	-4.79	4.91
p	<0.01	<0.01

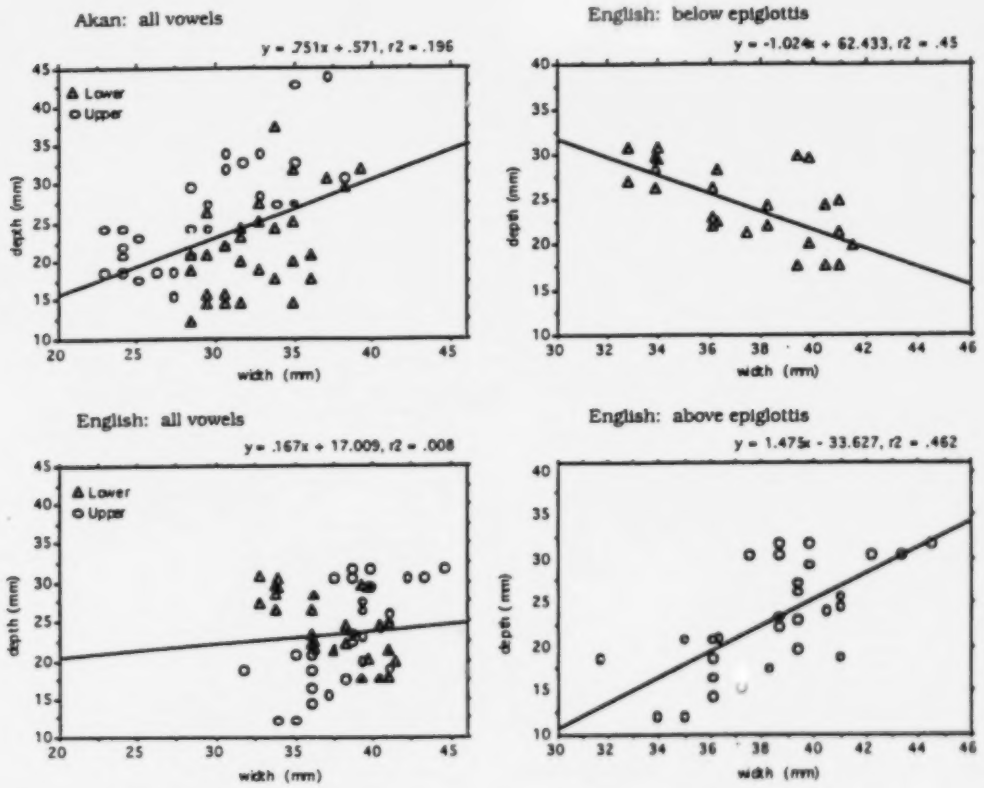


Figure 15. Regression of axial width × depth.

The fact that this divergence in behavior appears to occur exactly at the level of the base of the epiglottis suggests an anatomical explanation. It can be seen from the sagittal images that the base of the epiglottis is at approximately the same height as the hyoid, which serves as an anchor for the medial pharyngeal constrictor. The primary function of this muscle is to control pharyngeal aperture during swallowing in the region of the pharynx between the level of the hyoid and the upper larynx. Since this is where the Akan and English patterns diverge, the source of the cross-language difference in subepiglottal behavior (for these two subjects at least) may be the active involvement of this muscle in effecting the ATR contrast, and its lack of involvement in the Tense/Lax distinction: active control of pharyngeal aperture would account for the consistently larger (expanded - constricted) difference values observed for Akan, and its presumed absence would explain the small and inconsistent differences observed for English. The trading relation observed between English width and depth also supports the hypothesized non-involvement of the medial constrictor in the Tense/Lax mechanism: when GGP action on the tongue pulls it up and forward in Tense configurations, the relaxed constrictor does not oppose it, and so the lateral pharyngeal walls collapse slightly below the epiglottis as a consequence of the advanced tongue root. In Akan however active control of the constrictor results in the pharyngeal walls being tensed in +ATR configurations, preventing similar deformation of the lateral sides. Hardcastle (1976) refers to two modes of pharyngeal muscle contraction: an "isotonic" mode inducing sphincteral narrowing of pharyngeal aperture, and an "isometric" mode serving to tense pharyngeal walls without reducing aperture. Under this account of the differences in subepiglottal behavior between English and Akan, English does not actively control the pharyngeal constrictor, whereas Akan uses isotonic contraction of the pharyngeal constrictor to minimize pharyngeal aperture in [-ATR] configurations, and isometric contraction

in [+ATR] configurations to prevent deformation of pharyngeal lateral walls. Pharyngeal isometric tension may be significant in a different context: Hardcastle (1973) has suggested that it may be important in the production of tensed initial stops in Korean.

Assuming that the patterns observed for these two subjects are in fact representative of Akan and English, it is evident that the ATR and Tense/Lax distinctions are only superficially similar. In producing +ATR vowels Akan speakers enlarge the pharyngeal cavity by advancing the root of the tongue, lowering the larynx, and maintaining tension in the pharyngeal walls. In -ATR configurations the root is retracted, pharyngeal aperture is constricted, and the larynx is raised, minimizing pharyngeal volume. Because speakers appear to make adjustments to maintain relatively constant dorsum height across the two configurations, the ATR distinction is essentially a contrast in pharyngeal volume. Tense vowels in English are also articulated with an advanced tongue root, enlarging the pharyngeal cavity as a consequence; but below the level of the epiglottis this enlargement is counteracted by the deformation of the lateral walls from lack of constrictor tension. In lax configurations pharyngeal volume is smaller, but it is unclear whether this is due to actual constriction by the medial constrictor, or simply the relaxed position of the tongue root. The dorsum-raising effect of root advancement is not adjusted for in English, and instead constitutes an integral part of the contrast.

Generalizations of this sort are somewhat presumptuous given the limited scope of this study, and those made for English in particular should be viewed in the context of the Ladefoged *et al.* (1972) and Raphael & Bell-Berti (1975) findings mentioned above, showing that different English speakers produce the Tense/Lax contrast differently. Separate studies by Lindau in 1975 and 1979 involving four subjects each found consistent articulatory implementation of the ATR mechanism, so the generalizations made here for Akan are perhaps on firmer ground, especially as the sagittal results of this study replicated her findings. In any case, while different speakers of English may approximate more or less closely the Akan ATR articulatory mechanism, the point is that the Tense/Lax contrast is not identical to it, and must be represented by a different feature. Furthermore, because the Akan distinction appears to involve two muscles that the English contrast does not exploit (the medial pharyngeal constrictor and possibly the anterior genioglossus), from an articulatory standpoint it appears to be more complex with respect to the active control and coordination needed to produce it.

Conclusions

Although exploratory in scope, this study did succeed in achieving certain objectives. It demonstrated that despite its inherent drawbacks Magnetic Resonance Imaging can be a useful technique for obtaining information about static vocal tract configurations, one that will undoubtedly increase in importance as the technology is further refined. Measurements from the sagittal images collected here successfully replicated previous results from cineradiographic studies showing the significance of tongue root advancement and larynx height in effecting the Akan ATR contrast, and those obtained from axial images extended these results for the first time into the dimension of lateral width. The axial measurements confirm Lindau's position that it is the overall difference in pharyngeal volume that is relevant to the Akan vowel contrast, not just relative larynx or tongue root positions. Finally the parallel analysis of the English Tense/Lax contrast gave results showing that despite superficial similarities with the Akan ATR distinction, the two contrasts are not the same.

REFERENCES

Baer, T., Alfonso, P., & K. Honda (1988) Electromyography of the tongue muscles during vowels in /@pVp/ environment. *Annual Bulletin of the Research Institute for Logopedics and Phoniatrics (University of Tokyo)*, 22, 7-19.

Baer, T., Gore, J. C., Gracco, L. C., & Nye, P. W. (1991). Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels. *Journal of the Acoustical Society of America, 90 (2)*, 799-828.

Bradley, W. G., Newton, T. H., & Crooks, L. E. (1983). Physical principles of nuclear magnetic resonance. In T. H. Newton & D. G. Potts (Eds.), *Modern neuroradiology: Advanced imaging techniques* (pp. 15-62). San Francisco: Clavadel Press.

Clements, G. N. (1980). *Vowel harmony in a nonlinear generative phonology: An autosegmental model*. Indiana University Linguistics Club.

Dolphyne, F. A. (1988) *The Akan (Twi-Fante) Language*. Accra: Ghana Universities Press.

Halle, M., & Stevens, K. N. (1969). On the feature Advanced Tongue Root. *Quarterly Progress Report, 94*, Research Laboratory of Electronics, Massachusetts Institute of Technology, 209-215.

Hardcastle, W. J. (1973). Some observations on the tense-lax distinction in initial stops in Korean. *Journal of Phonetics, 1*, 263-272.

Hardcastle, W. J. (1976). *Physiology of speech production*. New York: Academic Press.

Harshman, R., Ladefoged, P., & Goldstein, L. (1977). Factor analysis of tongue shapes. *Journal of the Acoustical Society of America, 62*, 693-707.

Jackson, M. T. (1988). Phonetic theory and cross-linguistic variation in vowel articulation. *Working Papers in Phonetics, 71*, Los Angeles: University of California.

Ladefoged, P. (1964). *A phonetic study of West African languages*. Cambridge: University Press.

Ladefoged, P., DeClerk, J., Lindau, M., & Papcun, G. (1972). An auditory-motor theory of speech production. Working *Papers in Phonetics, 22*, 48-75. Los Angeles: University of California.

Lakshminarayanan, A. V., Lee, S., & McCutcheon, M. J. (1991). MR Imaging of the vocal tract during vowel production. *Journal of Magnetic Resonance Imaging, 1*, 71-76.

Lindau, M. (1975). [Features] For Vowels. *Working Papers in Phonetics, 30*. Los Angeles: University of California.

Lindau, M. (1979). The feature expanded. *Journal of Phonetics, 7*, 163-176.

Perkell, J. (1971). Physiology of speech production: a preliminary study of two suggested revisions of the features specifying vowels. *Quarterly Progress Report, 102*, Research Laboratory of Electronics, Massachusetts Institute of Technology, 123-139.

Raphael, L. J., & Bell-Berti, F. (1975). Tongue musculature and the feature of tension in English vowels. *Phonetica, 32*, 61-73.

Stewart, J. M. (1967). Tongue root position in Akan vowel harmony. *Phonetica, 16*, 185-204.

Stewart, J. M. (1971). Niger-Congo, Kwa. In T. Sebeok (Eds.), *Current trends in linguistics (pp.* 179-212). The Hague: Mouton.

FOOTNOTES

[1]High vowels from both harmony groups (and low /a/) also have nasalized counterparts which were not investigated in this study.

[2]In Clements' (1980) analysis of Akan vowel harmony /a/ is treated as an opaque vowel that induces the [-ATR] feature on any subsequent vowels. Some dialects of Akan have an additional ([+ATR]) front vowel /ɜ/ that harmonizes with /a/; however the exact distribution of this vowel is unclear, and it is ignored here.

[3]The magnetic resonance technique relies on strong magnetic fields to align the magnetic moments of hydrogen nuclei, and pulsed radio-frequency energy to set them into resonance. The signal intensity in a MR image depends on the density of the hydrogen nuclei in the scanned tissue and its resonance decay characteristics, determined by the atomic environment of the protons within each molecule. See Bradley *et al.* (1983) for further information.

[4]Personal communication from my Akan informant.

[5]To facilitate cross-language comparison in the following discussion I freely apply the terms "expanded" and "constricted" to both Akan and English, but do not mean to imply by this that the same feature mechanism is involved in both languages; nor am I referring to Lindau's term for the Akan feature. The terms are simply meant as shorthand physical descriptions of the respective contrasts, so that "expanded" for example should be understood as [+ATR] in the context of Akan, and [+Tense] in the context of English.

[6]Image version 1.29q by W. Rasband, NIH.

[7]Image quality was inferior to that obtained by Baer *et al.* (1991); however those researchers used custom-made cephalostats and longer scanning times in obtaining their best results.

[8]The scanning process resolved a given volume into a flat 256 x 256 pixel image. Sagittal scanning used a 280mm width x 280mm depth x 3mm thickness giving a resolution of 1.09mm/pix. Half the axial scans were done using a 280mm x 280mm x 5mm volume resulting in the same 1.09mm/pix resolution, and half were done using a 300mm x 300mm x 5mm volume giving a resolution of 1.17mm/pix. See Tables 1 and 2 for full scanning specifications used.

[9]Each volume element represents Σ (scan area * scan thickness), where scanned image thickness was a constant 5mm (see preceding footnote).

[10]The uppermost axial levels measured for Akan /e:ɛ/ show (slightly) larger depth values for the contracted variant, reversing the pattern found everywhere else. One possible explanation is that at these levels scanning has moved out of the region manipulated for the ATR contrast, and into the relatively invariant region of the tongue constriction characterizing the vowel.

[11]Both subjects reported feelings of nervousness when first placed inside the magnet; as the diameter of the central bore is approximately two feet only, claustrophobia is definitely a potential problem in studies of this type.

[12]For purposes of symmetry the uppermost level was dropped from the analysis, so that the five subepiglottal levels constituted the 'lower' group, and the epiglottal pivot and its four succeeding levels the 'upper.' ANOVA design (levels as repeated measures nested in group): level (measurement level, 1-5); group (below/above epiglottis); vowel (I/E/U); ATR (expanded/constricted); language (Akan/English). The level*vowel*ATR*language interaction was used as the error term (16df).

Haskins Laboratories Status Report on Speech Research
1993, SR-113, 131-134

Thai*

M. R. Kalaya Tingsabadh† and Arthur S. Abramson‡

Standard Thai is spoken by educated speakers in every part of Thailand, used in news broadcasts on radio and television, taught in school, and described in grammar books and dictionaries. It has evolved historically from Central Thai, the regional dialect of Bangkok and the surrounding provinces.

The transcription of a Thai translation of The North Wind and the Sun is based on recordings made by three cultivated speakers of the language, who were asked to read the passage in a relaxed way. In fact, we find them all to have used a fairly formal colloquial style, apparently equivalent to Eugénie J. A. Henderson's "combinative style" (Henderson, 1949). In a more deliberate reading of the text many words in the passage would be transcribed differently. The main features subject to such stylistic variation are vowel quantity, tone, and glottal stop. Thus, for example, /tɕ̀:/ 'but' is likely under weak stress to be /tɕ̀/ with a short vowel; the modal auxiliary /tɕà?/ 'about to' becomes /tɕā/, with change of tone from low to mid and loss of final glottal stop. The prosodic and syntactic factors that seem to be at work here remain to be thoroughly explored.

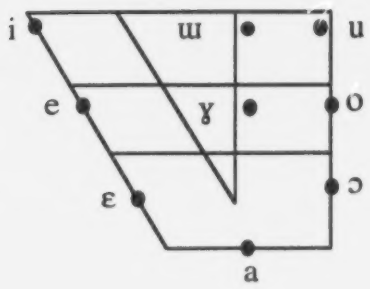
Consonants

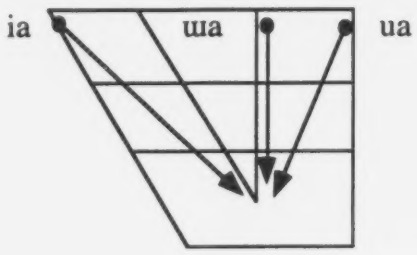
	Bilabial	Lab-dent.	Dental	Alveolar	Post-alveolar	Palatal	Velar	Glottal
Plosive	p pʰ b			t tʰ d			k kʰ	?
Nasal	m			n			ŋ	
Fricative		f		s				h
Affricate					tɕ tɕʰ			
Trill				r				
Approxi-mant						j	w	
Lateral Approx.				l				

p	pâ:	'elder aunt'	t	tām	'to pound'	k	kâ:ŋ	'fish bone'
pʰ	pʰâ:	'cloth'	tʰ	tʰām	'to do'	kʰ	kʰâ:ŋ	'side'
b	bā:	'insane'	d	dām	'black'	ŋ	ŋā:	'tusk'
m	mā:	'to come'	n	nā:	'ricefield'	w	wān	'day'
f	fāj	'pimple'	s	sǎj	'clear'	j	jā:m	'watchman'
			r	rák	'to love'	?	?ûan	'fat'
			l	lák	'to steal'	h	hǎj	'earthen jar'
			tɕ	tɕā:m	'to sneeze'			
			tɕʰ	tɕʰā:m	'dish'			

Vowels

There are nine vowels. Length is distinctive for all the vowels. (In some phonological treatments, /V:/ is analyzed as /VV/.) Although small spectral differences between short and long counterparts are psychoacoustically detectable and have some effect on vowel identification (Abramson & Ren, 1990), we find the differences too subtle to place with confidence in the vowel quadrilateral. The vowel /a/ in unstressed position, including the endings of the diphthongs /ia ɯa ua/, is likely to be somewhat raised in quality. The final segments of the other two sets of phonetic diphthongs: (1) [iu, eu, ɛu, ɔu, au, aːu, ɨau] and (2) [ai, aːi, ɔi, ɔːi, ui, ɤːi, uai, ɯai] are analyzed as /w/ and /j/ respectively.





i	krì̀t	'dagger'	iː	krì̀ːt	'to cut'	ia	rīān	'to study'
e	ʔēn	'ligament'	eː	ʔēːn	'to recline'	ɯa	rɯ̄ān	'house'
ɛ	pʰɛ́ʔ	'goat'	ɛː	pʰɛ́ːn	'to be defeated'	ua	ruān	'to be provocative'
a	fǎn	'to dream'	aː	fǎːn	'to slice'			
ɔ	klɔ̀ŋ	'box'	ɔː	klɔːŋ	'drum'			
o	kʰồn	'thick (soup)'	oː	kʰồːn	'to fell (a tree)'			
u	sù̀t	'last, rearmost'	uː	sù̀ːt	'to inhale'			
ɤ	ŋɤ̄n	'silver'	ɤː	dɤ̄ːn	'to walk'			
ɯ	kʰɯ̂n	'to go up'	ɯː	kʰɯ̂ɛn	'wave'			

Tones

There are five tones in Standard Thai: high /ˊ/, mid /¯/, low /ˋ/, rising /ˇ/, and falling /ˆ/.

kʰāː	'to get stuck'	kʰáː	'to engage in trade'
kʰàː	'galangal'		
kʰǎː	'I'	kʰâː	'leg'

Stress

Primary stress falls on the final syllable of a word. The last primary stress before the end of a major prosodic group commonly takes extra stress.

Conventions

The feature of aspiration is manifested in the expected fashion for the simple prevocalic oral stops /pʰ, tʰ, kʰ/. The fairly long lag between the release of the stop and the onset of voicing is filled with turbulence, i.e., noise-excitation of the relatively unimpeded supraglottal vocal tract. In the special case of the "aspirated" affricate /tɕʰ/, however, the noise during the voicing lag excites a narrow postalveolar constriction, thus giving rise to local turbulence. It is necessarily the case, then, that the constriction of the aspirated affricate lasts longer than that of the inaspirate (Abramson, 1989). Not surprisingly, it follows from these considerations that the aspiration of initial stops as the first element in clusters occurs during the articulation of the second element, which must be a member of the set /l, r, w/.

Only /p, t, k, ʔ, m, n, ŋ, w, j/ occur in syllable-final position. Final /p, t, k, ʔ/ have no audible release. The final oral plosives are said to be accompanied by simultaneous glottal closure (Henderson, 1964, Harris, 1972). Final /ʔ/ is omitted in unstressed positions. Initial /t/ and /f/ are velarized before close front vowels.

The consonant /r/ is realized most frequently as [ɾ] but also as [r]. Perceptual experiments (Abramson, 1962: 6–9) have shown that the distinction between /r/ and /l/ is not very robust; nevertheless, the normative attitude among speakers of Standard Thai is that they are separate phonemes, as given in Thai script. This distinction is rather well maintained by some cultivated speakers, especially in formal speech; however, many show much vacillation, with a tendency to favor the lateral phone [l] in the position of a single initial consonant. As the second element of initial consonant clusters, both /l/ and /r/ tend to be deleted altogether.

In plurisyllabic words, the low tone and the high tone on syllables containing the short vowel /a/ followed by the glottal stop in deliberate speech, become the mid tone when unstressed, with loss of the glottal stop.

Transcription of recorded passage

kʰāˈnàʔ tʰîˌlōmˈnɯ̌a lɛ́ːˌpʰráʔāˈtʰît | kāmˈlāŋ ˈtʰlāŋ kān ˈwâː | ˈkʰrāj tɕā ˈmîː pʰāˈlāŋ ˈmáːk kwā ˈkān | kɔ̂ ˈmîː ˈnákˌdɤ̄ːnˈtʰāːŋ ˈpʰûː ˈnɯ̀ŋ ˈdɤ̄ːn ˈpʰâːn ˈmāː | ˈsàjˈsɯ̌aˌkānˈnáːw ‖ ˌlōmˈnɯ̌a lɛ́ː ˌpʰráʔāˈtʰît ˈtɕɯ̄ŋ ˌtòkˈlōŋ kān ˈwâː | ˈkʰrāj tʰî ˌsǎːˈmáːt ˈtʰām hâj ˈnákˌdɤ̄ːnˈtʰāːŋ ˈpʰûː ˈnîː | ˈtʰɔ̀ːt ˈsɯ̌aˌkānˈnáːw ʔɔ̀k ˈdâj ˌsămˈrèt ˈkɔ̀ːn | tɕā ˈtʰɯ̄ː ˈwâː | pēn ˈpʰûː tʰî ˈmîː pʰāˈlāŋ ˈmáːk ˈkwâː ‖ lɛ́ʔ ˈlɛ́ːw | ˌlōmˈnɯ̌a kɔ̂ krāˈpʰɯ̄ː ˈpʰát ˈjàːŋ ˈsùt ˈrɛ̄ːŋ ‖ tɛ̀ ˈjīŋ ˈpʰát ˈrɛ̄ːŋ ˈmáːk ˈkʰûm ˈpʰīaŋ ˈdâj | ˈnákˌdɤ̄ːnˈtʰāːŋ kɔ̂ ˈjīŋ ˈduŋ ˈsɯ̌aˌkānˈnáːw ˈhâj krāˈtɕʰáp kàp ˈtūa ˈmáːk ˈkʰûm ˈpʰīaŋ ˈnán ‖ ˈlɛ́ʔ ˈnāj tʰî ˈsùt | ˌlōmˈnɯ̌a kɔ̂ ˈlɤ̄ːk ˈlóm ˈkʰwāːm pʰājāˈjáːm ‖ ˈtɕàːk ˈnán | ˌpʰráʔāˈtʰît tɕɯ̄ŋ ˈsàːt ˈsɛ̄ːŋ ʔān ˈrɔ̄ːn ˈrɛ̄ːŋ ʔɔ̀k ˈmāː ‖ ˈnákˌdɤ̄ːnˈtʰāːŋ kɔ̂ ˈtʰɔ̀ːt ˈsɯ̌aˌkānˈnáːw ʔɔ̀k ˈtʰān ˈtʰîː ‖ ˈnāj tʰî ˈsùt | ˌlōmˈnɯ̌a tɕɯ̄ŋ ˈtɕām ˈtɔ̄ŋ ˈjɔ̄ːm ˈráp ˈwâː | ˌpʰráʔāˈtʰît mîː pʰāˈlāŋ ˈmáːk ˈkwâː ˈtōn ‖

The passage in Thai script

ขณะที่ลมเหนือและพระอาทิตย์กำลังเถียงกันว่าใครจะมีพลังมากกว่ากัน ก็มีนักเดินทางผู้หนึ่งเดินผ่านมา ใส่เสื้อกันหนาว ลมเหนือและพระอาทิตย์จึงตกลงกันว่า ใครที่สามารถทำให้นักเดินทางผู้นี้ถอดเสื้อกันหนาวออกได้สำเร็จก่อน จะถือว่าเป็นผู้ที่มีพลังมากกว่า และแล้ว ลมเหนือก็กระพือพัดอย่างสุดแรง เต็มที่ พัดแรงมากขึ้นเพียงใด นักเดินทางก็ยิ่งดึงเสื้อกันหนาวให้กระชับกับตัวมากขึ้นเพียงนั้น และในที่สุดลมเหนือก็เลิกลมความพยายาม จาก นั้นพระอาทิตย์จึงส่องแสงอันร้อนแรงออกมา นักเดินทางก็ถอดเสื้อกันหนาวออกทันที ในที่สุดลมเหนือจึงจำต้องยอมรับว่าพระอาทิตย์มีพลังมากกว่าตน [1]

BEST COPY AVAILABLE

REFERENCES

Abramson, A. S. (1962). *The vowels and tones of Standard Thai: Acoustical measurements and experiments*. Bloomington: Indiana University Research Center in Anthropology, Folklore, and Linguistics, Publication 20.

Abramson, A. S. (1989). Laryngeal control in the plosives of Standard Thai. *Pasaa, 19*, 85–93.

Abramson, A. S., & Ren, N. (1990). Distinctive vowel length: Duration vs. spectrum in Thai. *Journal of Phonetics, 18*, 79–92.

Harris, J. G. (1972). Phonetic notes on some Siamese consonants. In J. G. Harris & R. B. Noss, (Eds.), *Tai phonetics and phonology* (pp. 8–22). Bangkok: Central Institute of English Language.

Henderson, E. J. A. (1949). Prosodies in Siamese: A study in synthesis. *Asia Major New Series 1*, 189–215.

Henderson, E. J. A. (1964). Marginalia to Siamese phonetic studies. In D. Abercrombie, D. B. Fry, P. A. D. MacCarthy, N. C. Scott, & J. L. M. Trim (Eds.), *In honour of Daniel Jones: Papers contributed on the occasion of his eightieth birthday 12 September 1961* (pp. 415–424). London: Longmans.

FOOTNOTES

*Journal of the International Phonetic Association, in press.

†Department of Linguistics, Faculty of Arts, Chulalongkorn University, Bangkok

‡Also Department of Linguistics, The University of Connecticut, Storrs

[1]We thank Dr. Chalida Rojanawathanavuthi, Dr. Kingkarn Thepkanjana, and Miss Surangkana Kaewnamdee, whose readings of the passage underlie our transcription. Part of the work of the second author was supported by Grant HD01994 from the U.S. National Institutes of Health to Haskins Laboratories.

Haskins Laboratories Status Report on Speech Research
1993, SR-113, 135-144

On the Relations between Learning to Spell and Learning to Read*

Donald Shankweiler† and Eric Lundquist‡

The study of spelling is oddly neglected by researchers in the cognitive sciences who devote themselves to reading. Experimentation and theories concerning printed word recognition continue to proliferate. Spelling, by contrast, has received short shrift, at least until fairly recently. It is apparent that in our preoccupation with reading, we have tended to downgrade spelling, passing it by as though it were a low-level skill learned chiefly by rote. However, a look beneath the surface at children's spellings quickly convinces one that the common assumption is false. The ability to spell is an achievement no less deserving of well- directed study than the ability to read. Yet spelling and reading are not quite opposite sides of a coin. Though each is party to a common code, the two skills are not identical. In view of this, it is important to discover how development of the ability to spell words is phased with development of skill in reading them, and to discover how each activity may influence the other. Thus, this chapter is concerned with the relationship between reading and writing.

It is appropriate to begin by asking what information an alphabetic orthography provides for a writer and reader, and to briefly review the possible reasons why beginners often find it difficult to understand the principle of alphabetic writing and to grasp how spellings represent linguistic structure. In this connection, would an orthography best suited for learning to spell differ from an orthography best suited for learning to read? The second section discusses how spelling and reading are interleaved in a child newly introduced to the orthography of English. Here, one central question is precedence: Does the ability to read words precede the ability to spell them, or, alternatively, might some children be ready to apply the alphabetic principle in writing before they can do so in reading? A related question is strategy. Do children sometimes approach the two tasks in very different ways? Finally, the last section discusses how analysis of children's spellings may illuminate aspects of orthographic learning that are not readily accessible in the study of reading.

We are indebted to Professor Peter Bryant of the University of Oxford for making it possible for one of us (DS) to study a group of children whose data are discussed in this chapter, and for help with the analysis and much valuable discussion. Thanks are due to Stephen Crain, Anne Fowler, Ram Frost, Leonard Katz, Alvin Liberman, and Hyla Rubin for their comments on earlier drafts. This work was supported in part by Grant HD-01994 to Haskins Laboratories from the National Institute of Child Health and Human Development.

HOW WRITERS AND READERS ARE EQUIPPED TO COPE WITH THE INFORMATION PROVIDED BY AN ALPHABETIC SYSTEM

Writing differs from natural and conventional signs in that it represents linguistic units, not meanings directly (DeFrancis, 1989; Mattingly, 1992). The question of how the orthography maps the language is centrally relevant to the course of acquisition of reading and spelling. All forms of writing permit the reader to recover the individual words of a linguistic message. Given that representation of words is the essence of writing, it is important to appreciate that words are phonological structures. To apprehend a word, whether in speech or in print, is thus to apprehend (among other things) its phonology. But in the manner of doing this, A. M. Liberman (1989; 1992) notes that there is a fundamental difference between speech on the one hand and reading and writing on the other. For a speaker or listener who knows a language, the language apparatus produces and retrieves phonological structures by means of processes that function automatically below the conscious level. Thus,

Liberman notes that to utter a word one does not need to know how the word is spelled, or even that it can be spelled. The speech apparatus that forms part of the species-specific biological specialization for language "spells" the word for the speaker (that is, it identifies and orders the segments). In contrast, writing a word, or reading one, brings to the fore the need for some *explicit* understanding of the word's internal structure. Since in an alphabetic system, it is primarily phonemes that are mapped, those who succeed in mastering the system would therefore need to grasp the phonemic principle and be able to analyze words as sequences of phonemes.

The need that alphabetic orthographies present for conscious apprehension of phonemic structure poses special difficulties for a beginner (see Gleitman & Rozin, 1977; I. Y. Liberman, 1973; Liberman, Shankweiler, Fischer, & Carter, 1974). The nub of the problem is this: phonemes are an abstraction from speech, they are not speech sounds as such. Hence, the nature of the relation between alphabetic writing and speech is necessarily indirect and, as we now know, often proves difficult for a child or a beginner of any age to apprehend. In order to understand why this is so it will pay us to dwell for a moment on the ways in which it is misleading to suppose that an alphabetic orthography represents speech sounds (see Liberman, Rubin, Duques, & Carlisle, 1985; Liberman et al., 1974).

First, the letters do not stand for segments that are acoustically isolable in the speech signal. So, for example, one does not find consonants and vowels neatly segmented in a spectrogram in correspondence with the way they are represented in print. Instead phonemes are co-articulated, thus overlappingly produced, in syllable- sized bundles. Accordingly, apprehension of the separate existences of phonemes and their serial order requires that one adopt an analytic stance that differs from the stance we ordinarily adopt in speech communications, in which the attention is directed to the content of an utterance, not to its phonological form. In view of this, it is not surprising to discover that preschool children have difficulty in segmenting spoken words by phoneme (see Liberman et al., 1989; Morais, 1991 for reviews).

Without some awareness of phonemic segmentation, it would be impossible for a beginning reader or writer to make sense of the match between the structure of the printed word and the structure of the spoken word. So, for example, writers and readers can take advantage of the fact that the printed word CLAP has four segments only if they are aware that the spoken word "clap" has four (phonemic) segments. Accordingly, in order to master an alphabetic system it is not enough to know the phonetic values of the letters. That knowledge, necessary though it is, is not sufficient. In order to fully grasp the alphabetic principle, it is necessary, in addition, to have the ability to decompose spoken words phonemically. Indeed, experience shows that there are many children who know letter-phoneme correspondences yet have poor word decoding skills (Liberman, 1971).

Considerable evidence now exists that children's skill in segmenting words phonemically and their progress in reading are, in fact, causally linked (e.g., Adams, 1990; Ball & Blachman, 1988; 1991; Bradley & Bryant, 1983; Byrne & Fielding-Barnsley, 1991; Goswami & Bryant, 1990; Lundberg, Frost, & Petersen, 1988). One would also expect to find that the same kind of relationship prevails between phoneme segmentation abilities and spelling. And, indeed, the data are consistent with that expectation. Studies by Zifcak (1984) and Liberman et al. (1985) have shown substantial correlations between performance on tests of phoneme segmentation of spoken words and the degree to which all the phonemes are represented in children's spellings. The findings of Rohl and Tunmer (1988) confirm this association. They compared matched groups of older poor spellers with younger normal ones and found that the poor spellers did significantly less well on a test of phoneme segmentation. (See also Bruck & Treiman, 1990; Juel, Griffith, & Gough, 1986, and Perin, 1983).

The complex relation between phonemic segments and the physical topography of speech is one sense in which alphabetic writing represents speech sounds only remotely. This, we have supposed, constitutes an obstacle for the beginning reader/writer to the extent that it makes the alphabetic principle difficult to grasp and difficult to apply. Two further sources of the abstractness of the orthography should also be mentioned, which may be especially relevant to the later stages of learning to read and to spell.

First, alphabetic orthographies are selective in regard to those aspects of phonological structure that receive explicit representation in the spellings of words (Klima, 1972; Liberman et al., 1985). No natural writing system incorporates the kind of phonetic detail that is captured in the special-purpose phonetic writing that linguists

use. Much context- conditioned phonetic variation is ignored in conventional alphabetic writing,[1] in addition to the variation associated with dialect and idiolect. Hence, conventional writing does not aim to capture the phonetic surface of speech, but aims instead to create a more generally useful abstraction. It is enlightening to note, in this connection, that young children's "invented spellings"[2] often differ from the standard system in treating English writing as though it were more nearly phonetic than it is (Read, 1971; 1986).

A second source of abstractness stems from the fact that the spelling of English is more nearly morphophonemic than phonemic. English orthography gives greater weight to the morphological structure of words than is the case with some other alphabetic orthographies, for example, Italian (see Cossu, Shankweiler, Liberman, Katz & Tola, 1988) and Serbo-Croatian (see Ognjenović, Lukatela, Feldman, & Turvey, 1983). Examples of morphological penetration in the writing of English words are easy to find. A ubiquitous phenomenon is the consistent use of s to mark the plural morpheme, even in those words, like DOGS, in which the suffix is pronounced not [s], but [z]. The morphemic aspect of English writing appears also in spellings that distinguish words that are homophones, for example, CITE, SITE; RIGHT, WRITE.

The knowledge that spellings of some English words may sacrifice phonological transparency to capture morphological relationships brings into perspective certain seeming irregularities, as several writers have noted (Chomsky & Halle, 1968; Klima, 1972; Venezky, 1970). Homophone spellings are instances in which the two modes of representation, the phonemic and the morphemic, are partially in conflict (DeFrancis, 1989). In these spellings the principle of alphabetic writing is compromised to a degree, but it is not abandoned, since most letters are typically shared between words that have a common pronunciation. A lexical distinction in homophone pairs is ordinarily indicated by the change of only a letter or two. Thus, homophone spellings in English present an irregularity from a narrowly phonological standpoint, while nonetheless keeping the irregularity within circumscribed limits.

Such examples are telling. They led DeFrancis (1989) to make a novel and stimulating suggestion: that the needs of readers and writers may actually conflict to some degree. The convention of distinct spellings for homophones would benefit readers by removing lexical ambiguity in cases in which context does not immediately resolve the matter. Writers, on the other hand, would perhaps be better served by a system that minimizes inconsistencies in mapping the surface phonology. For writers, the presence of homophones which are distinguished by their spellings increases the arbitrariness of the orthography, and hence the burden on memory. Because it has to serve for both purposes, the standard system can be regarded as a compromise, in some instances favoring readers and in other instances favoring writers.

Scrutiny of the words that users of English find difficult to spell confirms that morphologically complex words are among those most often misspelled (Carlisle, 1987; Fischer, Shankweiler, & Liberman, 1985). Carlisle (1988) notes that in derived words the attachment of a suffix to the base may involve a simple addition resulting in no change in either pronunciation or spelling of the base (ENJOY, ENJOYMENT). Alternatively, the addition may result in a pronunciation change in the base (HEAL, HEALTH), a spelling change but not a pronunciation change (GLORY, GLORIOUS) or a case in which both spelling and pronunciation change (DEEP, DEPTH). Difficulties in spelling morphologically complex words appear to stem in part from their phonological complexity and irregular spellings. But they may also stem from failure to recognize and accurately partition derivationally related words. Carlisle (1988) tested school children aged 8 to 13 for morphological awareness. They were asked to respond orally with the appropriate derived form, given the base followed by a cueing sentence designed to prompt a derivative word (e.g., "Magic. The show was performed by a _____"). It was found that awareness of derivative relationships was very limited in the youngest children, especially in cases in which the base undergoes phonological change in the derived form (as in the above example). Moreover, the ability to produce derived forms has proven deficient in children and adults who are poor spellers (Carlisle, 1987; Rubin, 1988). All in all, the evidence supports the expectation that both phonologic and morphologic aspects of linguistic awareness are relevant to success in spelling and reading.

So far we have discussed the common basis of reading and writing, pointing first to the great divide that separates speech processes on the one hand from orthographic processes on the other. Then we proceeded to identify the factors that make learning an alphabetic system difficult. The

idea was also introduced that reading and spelling may tax orthographic knowledge in somewhat different ways. It is to these differences that we turn next.

CAN CHILDREN APPLY THE ALPHABETIC PRINCIPLE IN SPELLING BEFORE THEY ARE ABLE TO APPLY IT IN READING?

The possibility that the needs of readers and writers may differ with respect to the kind of orthographic mapping that is easiest to learn raises the broader issue of the relation between learning to write and learning to read. Does one precede the other? Do children adopt different strategies for the one than for the other? To answer these questions we will want to examine what is known about how spelling articulates with reading in new learners.

As to the first question, one may wonder whether precedence is really an issue. Just as in primary language development, where it is often noted that children's perceptual skills run ahead of their skills in production, so in written language, too, it would seem commonsensical to suppose that a new learner's ability to read words would exceed the ability to spell them. Most users of English orthography have probably had the experience of being unsure how to spell some words that they recognize reliably in reading. Contributing to the difficulty is the fact that there is usually more than one way for a word to be spelled that would equivalently represent its phonological structure. (Consider, for example, "clene" and "cleen" as equivalent transcriptions of the word *clean*). The reader's task is to recognize the correspondence between a letter string that stands for a word (i.e., its morphophonological structure) and the corresponding word in the lexicon. It is not required that the reader know exactly how to spell a word in order to read it—only that the printed form (together with the context) should provide sufficient cues to prompt recognition of the represented word and not some other word. In contrast, the writer must generate the one (and ordinarily only one) spelling that corresponds to the conventional standard. So it is natural to assume that spelling words requires greater orthographic knowledge than reading them. We therefore might expect that a beginner would have the ability to read many words before necessarily being able to spell them correctly.

Nonetheless, questions about precedence in the development of reading and writing have arisen repeatedly. Some writers have suggested that, contrary to the view that reading is easier, children may indeed be ready to write words, in some fashion, before they are able to use the alphabetic principle productively in reading. Montessori (1964) expressed this view, and it has more recently been articulated by several prominent researchers. In part, these claims are based on experiences with preschool children who were already writing using their own invented spellings. Carol Chomsky (1971; 1979) stressed that many young writers do this at a time when they cannot read, and, indeed, may show little interest in reading what they have written. Others who have proposed a lack of coordination between spelling and reading in children's acquisition of literacy are Bradley and Bryant (1979), Frith (1980), and Goswami and Bryant (1990).

In order to discuss the question of precedence we must first consider how we are going to define spelling and reading. By spelling, do we mean spelling a word according to conventional spelling? To adopt that criterion would ignore the phenomenon of children's invented spelling. That would seem unwise since it is well-established that some children are able to write more or less phonologically before they know standard spellings (Read, 1971; 1986). It would be appropriate for some purposes to credit a child for spelling a word if the spelling the child produces approximates the word closely enough that it can be read as the intended word.

The criterion of reading is in one sense less problematical, but in another sense it is more so. For someone to be said to have read a word, that word, and not some other word (or nonword) must have been produced in response to the printed form. It is also relevant to ask how the response was arrived at. Words written in an alphabetic system can be approached in a phonologically analytic fashion or, alternatively, they can be learned and remembered holistically (i.e., as though they were logographs). As Gough and Hillinger (1980) stress, the difficulty with the logographic strategy is that it is self-limiting because it does not enable a reader to read new words. Moreover, as the vocabulary grows and the number of visually similar words increases, the memory burden becomes severe and the logographic strategy becomes progressively more inaccurate. Should we therefore consider someone a reader if she can identify high frequency words, but cannot read low frequency words or nonwords? There is some consensus that we should not (e.g.,

Adams, 1990; Gleitman & Rozin, 1977; Gough & Hillinger, 1980; Liberman & Shankweiler, 1979). The possibility of reading new words, not previously encountered in print, is a special advantage conferred by an alphabetic system. It is reasonable to suppose that someone who has mastered the system will possess that ability.

However, in the view of some students of reading, most children when they begin to read, and perhaps for a considerable time afterward, read logographically, and only later learn to exploit the alphabetic principle (Bradley & Bryant, 1979; Byrne, 1992; Gough & Hillinger, 1980). Given the absence of agreement as to what is to be taken as sufficient evidence of reading ability, the question of whether spelling or reading comes first is less the issue than whether children initially employ discrepant strategies for reading and writing.

The strategy question is brought into focus by Goswami and Bryant (1990). As noted above, they suppose that the child's initial strategy in reading (the default strategy) is to approach alphabetically written words as though they were logographs. They contend that children tend to do this even when they have had instruction designed to promote phonemic awareness. Reading analytically might require more advanced word analysis skills than are available to most beginning readers. Writing, on the other hand, forces the child to think in terms of segments. The process of alphabetic writing is by its nature segmental and sequential: The writer forms one letter at a time and must order the letters according to some plan. Thus, Goswami and Bryant suppose that children's initial approaches to writing would tend to be phonologically analytic. Goswami and Bryant (1990) find it paradoxical that children's newly found phonological awareness, which most often is introduced in the context of instruction in reading, has an immediate effect on their spelling, but not on their reading. "So at first there is a discrepancy and a separation between children's reading and spelling. It is still not clear why children are so willing to break up words into phonemes when they write, and yet are so reluctant to think in terms of phonemes when they read (p. 148)."

Bryant and his colleagues (see especially Bradley and Bryant, 1979) deserve much credit for grasping the need for a coordinated approach to the study of reading and spelling. They recognized that this undertaking would require testing children on reading and spelling the same words. It is well known that performance on reading and spelling tests are highly correlated, at least in older children and adults (Perfetti, 1985; Shankweiler & Liberman, 1972). Bradley and Bryant stressed that the correlation between reading and spelling scores depends on the words chosen. They proposed that the words that children at the beginning stages find difficult to read are not always the words that are difficult to spell, and vice versa. Words that tended to be read correctly but misspelled were words whose spellings presented some irregularity, like EGG or LIGHT, whereas words spelled and not read tended to be regular words, like MAT and BUN (Bradley & Bryant, 1979).

The finding that the spell-only words and the read-only words did not overlap very much in the beginning would lend support to the hypothesis that children at this stage use different strategies for spelling and reading. The greater difficulty in spelling irregular words is what one would expect if the children were attempting to spell according to regular letter-to-phoneme correspondences. They would tend to regularize the irregular words and thus get them wrong. Moreover, the failure to read regular words suggests that the children were using some nonanalytic strategy for reading, responding perhaps to visual similarity. That would make them prone to miss easy words whenever their appearance is confusable with other words that look similar. If they were reading analytically they would read these words correctly. Thus, Bryant and his colleagues cite findings that seem to underscore the differences between early reading and spelling.

Should we, then, accept Goswami and Bryant's paradox and suppose that reading and writing are cognitively disjunct at the early stages, even in children who have received training in phonological awareness? We think not. First, as the succeeding section shows, some data (to which we turn next) point to concurrent development of reading and spelling skills. Secondly, it is too early to assess fully the impact on children's reading and spelling of the several experimental approaches to instruction in phonological awareness (e.g., Ball & Blachman, 1988; 1991; Blachman, 1991; Byrne & Fielding-Barnsley, 1991; in press). Therefore, we believe that the question must remain open.

A new research study, which coordinated the investigation of spelling and reading in six year olds (the subjects were selected only for age), does not find evidence that incompatible strategies are employed by beginners (Shankweiler, 1992). Unlike the Bradley and Bryant study, the test

words in this experiment included no words with irregular spellings. The test words did contain phonological complexities, however. Each contained a consonant cluster at the beginning or the end.

There was a wide range in level of achievement within this group of six year olds. Nine of the 26 children were unable to read and spell more than one word correctly. The remaining 17 were able to read a mean of 70 percent of the words correctly but were able to correctly spell only 39 percent. These findings show that the spelling difficulties of beginners are not confined to irregular words.[3] Regularly spelled words can cause difficulty if they are phonologically complex, as when they contain consonant clusters. With the exception of one child, all read more words correctly than they were able to spell. Finally, analytic skill in reading, as indexed by ability to read nonwords, was almost perfectly correlated ($r = .93$) with spelling performance (on a variety of real words).[4] These data do not sit well with the conclusion that early reading and spelling are cognitively dissociated. On the contrary, the findings lend support to the idea that skill in reading and spelling tend to develop concurrently over a wide range of individual differences in attainment.

It is notable that spelling accuracy consistently lagged somewhat behind reading. Only 6 percent of the words were spelled correctly and read incorrectly, whereas 37 percent were read and not spelled. Thus the children showed what might be expected to be true generally: that spelling the words would prove to be more difficult than reading them, if by reading we mean correct identification of individual words, and by spelling we mean spelling these words according to standard conventions.

INTERPRETING ERROR PATTERNS IN SPELLING AND READING

So far we have been comparing spelling and reading at a coarse level of analysis. To address more rigorously the question of whether new learners use similar or dissimilar strategies for spelling and reading we would wish to make a detailed comparison between the error pattern in spelling words and reading them. But, as it happens, this turns out to be a difficult thing to do.

Problems of comparability

Most of the published information on the correlations between reading and spelling scores is based simply on right/wrong scoring. This approach has the disadvantage of throwing away much of the potential information in the incorrect responses. It fails to distinguish reading errors that are near misses from errors that are wild guesses, and it does not distinguish misspellings that capture much of a word's phonological structure from those that capture little of it. If we give partial credit for wrong responses, we must create a scheme to evaluate the many possible ways of misspelling a word and assign relative weights to each.

As an illustration of how we might proceed, we turn again to the research study last described (Shankweiler, 1992). In this study, reading was assessed by the Decoding Skills Test (DST, Richardson, & Di Benedetto, 1986). The test consists of 60 real words, chosen to give representation to the major spelling patterns of English, and, importantly, it also includes an equal number of matching nonwords, the latter formed by changing one to three letters in each of the corresponding words. For the purposes at hand, phonotactically legal nonwords constitute the best measure of reading for assessing the skills of the beginning readerbecause only these can provide a true measure of decoding skill. Because they are truly unfamiliar entities, nonwords test whether a reader's knowledge of the orthography is productive. As noted earlier, only that kind of knowledge enables someone to read new words not previously encountered in print (see Shankweiler, Crain, Brady, & Macaruso, 1992). Responses to the Decoding Skills Test were recorded on audiotape and transcribed in IPA phonemic symbols for later comparison with the spelling measures.

To gain a fine-grained measure of spelling for comparison with the reading error measures, the children's written spellings were scored phoneme by phoneme, using the following categories:

Correct spelling
Phonologically acceptable substitute
(e.g., k for ck)
Phonologically unacceptable substitute
(e.g., c for ch)
Phoneme not represented

When we try to compare the error pattern in reading and spelling, we encounter a further difficulty: Reading is a covert process that is assessed only by its effects. One cannot directly infer what goes on in the head when someone attempts to read a word. When we ask the child to read aloud unconnected words in list form, we

encounter an obstacle: children are often unwilling to make their guesses public. Of course, a beginning reader who is stuck on a particular word may be entertaining a specific hypothesis about the word's identity, but in the absence of an overt response, we cannot discover the hypothesis and use it as a basis for inferring the source of the difficulty.

Writing, on the other hand, leaves a visible record of the writer's hypothesis about how to spell a word. The findings of the study we have been discussing bear this out. Many of the children declined the experimenter's invitation to guess at the words they were having difficulty in reading. Yet the same children produced a spelling for nearly every word they were asked to write. The upshot is that we have nearly a complete set of responses to the spelling test, but many gaps in the record occur on the corresponding items on the reading test. This yields an unsatisfactory data base for comparing the error pattern in spelling and reading. Thus, the kind of word-by-word comparison we would like to make may be unattainable.

Nonetheless, there is much to be gained by a linguistic analysis of children's spellings. Indeed, it is chiefly through their writing, and not through their reading, that children reveal their hypotheses about the infrastructure of words.

Children's conceptions of the infrastructure of words as revealed in their spellings

When encouraged to invent spellings for words, young children invent a system that is more compatible with their linguistic intuitions than the standard system. Whether the result corresponds to standard form is simply not a question that would occur to the child at this stage. In Carol Chomsky's words, creative spellers "appear to be more interested in the activity than the product (1979, p. 46)." There is evidence that children's invented spellings tend to be closer to the phonetic surface than the spellings of the standard system (Read, 1986). The standard system of English, as we noted, maps lexical items at a level that is highly abstract, both because the conventional system is morphophonemic, and because it tends not to transcribe phonetic detail that is predictable from general phonological rules.

In the comparative study of reading and writing in six year olds which we have discussed (Shankweiler, 1992), even the least-advanced beginners, who wrote only a single letter to represent an entire word, usually chose a consonant that could represent the first phoneme in the word. A child who does this is apparently aware that letters represent phonological entities even though she is not yet able to analyze the internal structure of the syllable. Altogether, first consonants were represented in 95% of cases. There was a strong tendency to omit the second segment of a consonant cluster: that is, the L in CL, the T in ST, the M in SM, the R in CR, and so forth. These were omitted in 56% of occurrences, yet when these consonants occurred alone in initial position, they were rarely omitted. Bruck and Treiman (1990) report the same trends, both in normal children and dyslexics. The tendency to omit the second segment from an initial cluster fits with Treiman's idea (1992) that children may initially use letters to represent syllable onsets and rimes rather than phonemes.[5]

The ability to represent the second segment of initial consonant clusters was a very good predictor of overall spelling achievement. It was also a good predictor of the accuracy of word reading. Regression analysis showed that this part score accounted for 45 percent of the variance in either spelling or reading when a different set of words is tested, after age, vocabulary (Dunn, Dunn, & Whetton, 1982) and a measure of phonemic segmentation skill (Kirtley, 1989) had already been entered. Representation of the interior segment in final clusters does almost as well when entered in the regression. The results of fine scoring give further support to the view that reading and spelling skill are closely linked even in beginners.

Why are consonant clusters a special source of difficulty? Two possibilities might be considered, each related to the phonetic complexity of clusters. First, it is well known that clusters cause pronunciation difficulties for young children. Perhaps the spelling error signals a general tendency to simplify these consonant clusters - a failure to perceive and produce them as two phonemes. But there was no indication that this was the case. All the children could pronounce the cluster words without difficulty.

An alternative possibility is that the children had difficulty in conceptually breaking clusters apart and representing them as two phonemes. In that case, the difficulty in spelling could be seen as a problem in phonological awareness. So, also, could the problems in reading the cluster words. Reading analytically would require the reader to decompose the word into its constituent segments, and the presence of clusters would increase the difficulty of making this analysis.

Research conducted during the past two decades has shown that phonological awareness is not all of a piece. Full phoneme awareness is a late stage in a process of maturation and learning that takes years to complete (Bradley & Bryant, 1983; Liberman et al., 1974; Morais, Cary, Alegria, & Bertelson, 1979; Treiman & Zukowski, 1991). Although the order of acquisition is not completely settled, there is evidence that before they can segment by phoneme children are able to segment spoken words using larger sublexical units—onsets and rimes, and syllables, particularly stressed syllables that rhyme (Brady, Gipstein, & Fowler, 1992; Liberman et al., 1974; Treiman, 1992).

The role of literacy instruction in fostering the development of phonological awareness has been much discussed in the research literature (See chapters in Brady & Shankweiler, 1991, and in Gough, Ehri, & Treiman, 1992). In this connection, Treiman (1991) urges that an analysis of spelling is the best route by which to study those aspects of phonological awareness that depend on experience with reading and writing. We would tend to agree. This is not to say, however, that *writing, but not reading* would feed this development in young children. It is to be expected that a child's interest and curiosity about the one activity would encourage and nourish an interest in the other.[6]

To sum up, because reading and writing are secondary language functions derived from spoken language, they display a very different course of acquisition than speech itself: unlike speech, mastery of alphabetic writing requires facility in decomposing words into phonemes and morphemes. Since both reading and writing depend upon grasp of the alphabetic principle, it could be expected that both would develop concurrently, though spelling, being the more difficult, would progress more slowly. Several researchers, however, have raised challenging questions about the order of precedence, suggesting that spelling, due to the inherently segmental nature of writing words alphabetically, emerges earlier than the ability to decode in reading. At present, the evidence is mixed. It is significant that recent research comparing children's reading and spelling errors indicates that in both spelling and reading, regularly spelled words present difficulties to beginners when the words contain phonologically-complex consonant clusters. Thus, beginners' difficulties in reading and spelling do not necessarily involve different kinds of words, as had been suggested earlier. This undercuts the claim of incompatible strategies.

Whether a child initially adopts a logographic or an analytic strategy for reading may depend in large part on the kind of pre-reading instruction the child was provided with. There is evidence that both phonological awareness and knowledge of letter-phoneme correspondences are important to promote grasp of the alphabetic principle, and are thus important to skill in spelling and decoding (Ball & Blachman, 1988; 1991; Bradley & Bryant, 1983; Byrne & Fielding-Barnsley, 1991; Gough, Juel & Griffith, 1992). Neither is sufficient alone. The phasing of these two necessary components of instruction may turn out to be critical in determining the child's initial approach to the orthography.

REFERENCES

Adams, M. J. (1990). Beginning to read: Thinking and learning about print. Cambridge, MA: MIT Press.

Ball, E. W., & Blachman, B. A. (1988). Phoneme segmentation training: Effect on reading readiness. Annals of Dyslexia, 38, 208-225.

Ball, E. W., & Blachman, B. A. (1991). Does phoneme awareness training in kindergarten make a difference in early word recognition and developmental spelling? Reading Research Quarterly, 26, 49-66.

Blachman, B. A. (1991). Phonological awareness: Implications for prereading and early reading instruction. In S. A. Brady & D. P. Shankweiler (Eds.), Phonological processes in literacy: A tribute to Isabelle Y. Liberman. Hillsdale, NJ: Lawrence Erlbaum Associates.

Bradley, L., & Bryant, P. E. (1979). The independence of reading and spelling in backward and normal readers. Developmental Medicine and Child Neurology, 21, 504-514.

Bradley, L., & Bryant, P. E. (1983). Categorizing sounds and learning to read—a causal connection. Nature, 30, 419-421.

Brady, S. A., Gipstein, M., & Fowler, A. E. (1992). The development of phonological awareness in preschoolers. Presentation at AERA annual meeting, April 1992.

Brady, S. A., & Shankweiler, D. P. (1991). *Phonological processes in literacy: A tribute to Isabelle Y. Liberman.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Bruck, M., & Treiman, R. (1990). Phonological awareness and spelling in normal children and dyslexics: The case of initial consonant clusters. *Journal of Experimental Child Psychology, 50,* 156-178.

Byrne, B. (1992). Studies in the Acquisition Procedure for Reading: Rationale, hypotheses, and data. In P. B. Gough, L. C. Ehri, & R. Treiman (Eds.), *Reading acquisition* (pp. 1-34). Hillsdale, NJ: Lawrence Erlbaum Associates.

Byrne, B., & Fielding-Barnsley, R. (1990). Acquiring the alphabetic principle: A case for teaching recognition of phoneme identity. *Journal of Educational Psychology, 82,* 805-812.

Byrne, B., & Fielding-Barnsley, R. (1991). Evaluation of a program to teach phonemic awareness to young children. *Journal of Educational Psychology, 83,* 451-455.

Carlisle, J. F. (1987). The use of morphological knowledge in spelling derived forms by Learning Disabled and Normal Students. *Annals of Dyslexia, 37,* 90-108

Carlisle, J. F. (1988). Knowledge of derivational morphology and spelling ability in fourth, sixth, and eighth graders. *Applied Psycholinguistics, 9,* 247-266.

Chomsky, C. (1971). Write first, read later. *Childhood Education, 47,* 296-299.

Chomsky, C. (1979). Approaching reading through invented spelling. In L. B. Resnick & P. A. Weaver (Eds.), *Theory and practice of early reading, vol. 2* (pp. 43-65). Hillsdale, NJ: Lawrence Erlbaum Associates.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English.* New York: Harper and Row.

Cossu, G., Shankweiler, D., Liberman, I. Y., Tola, G., & Katz, L. (1988). Phoneme and syllable awareness in Italian children. *Applied Psycholinguistics, 9,* 1-16.

DeFrancis, J. (1989). *Visible speech: The diverse oneness of writing systems.* Honolulu: University of Hawaii Press.

Dunn, L. M., Dunn, L. M., & Whetton, C. (1982). *British picture vocabulary scale.* Berks., England: NFER-Nelson.

Ehri, L. C. (1989). The development of spelling knowledge and its role in reading acquisition and reading disability. *Journal of Learning Disabilities, 22,* 356-365.

Ehri, L. C., & Wilce, L. S. (1987). Does learning to spell help beginners learn to read words? *Reading Research Quarterly, 22,* 47-64.

Fischer, F. W., Shankweiler, D., & Liberman, I. Y. (1985). Spelling proficiency and sensitivity to word structure. *Journal of Memory and Language, 24,* 423-441.

Frith, U. (1980). Unexpected spelling problems. In U. Frith (Ed.), *Cognitive processes in spelling.* London: Academic Press.

Gleitman, L. R., & Rozin, P. (1977). The structure and acquisition of reading I: Relations between orthographies and the structure of the language. In A. S. Reber & D. L. Scarborough (Eds.), *Toward a psychology of reading.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Goswami, U., & Bryant, P. (1990). *Phonological skills and learning to read.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Gough, P. B., Ehri, L. C., & Treiman, R. (1992). *Reading acquisition.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Gough, P. B., & Hillinger, M. L. (1980). Learning to read: An unnatural act. *Bulletin of the Orton Dyslexia Society, 30,* 179-196.

Gough, P. B., Juel, C., & Griffith, P. L. (1992). Reading, Spelling, and the orthographic cipher. In P. B. Gough, L. C. Ehri, & R. Treiman (Eds.), *Reading acquisition* (pp. 35-48). Hillsdale, NJ: Lawrence Erlbaum Associates.

Juel, C., Griffith, P. L., & Gough, P. B. (1986). Acquisition of literacy: A longitudinal study of children in first and second grade. *Journal of Educational Psychology, 78,* 243-255.

Kirtley, C. L. M. (1989). Onset and rime in children's phonological development. Unpublished doctoral dissertation, University of Oxford.

Klima, E. S. (1972). How alphabets might reflect language. In J. F. Kavanagh & I. G. Mattingly (Eds.), *Language by ear and by eye.* Cambridge, MA: MIT Press.

Liberman, A. M. (1989). Reading is hard just because listening is easy. In C. von Euler, I. Lundberg & G. Lennerstrand (Eds.), *Wenner-Gren Symposium Series 54, Brain and Reading.* London: Macmillan.

Liberman, A. M. (1992). The relation of speech to reading and writing. In L. Katz & R. Frost (Eds.), Orthography, phonology, morphology, and meaning (pp. 167-178). Amsterdam: Elsevier Science Publishers.

Liberman, I. Y. (1971). Basic research in speech and lateralization of language: Some implications for reading disability. *Bulletin of the Orton Society, 21,* 71-87.

Liberman, I. Y. (1973). Segmentation of the spoken word. *Bulletin of the Orton Society, 23,* 65-77.

Liberman, I. Y., Rubin, H., Duques, S., & Carlisle, J. (1985). Linguistic abilities and spelling proficiency in kindergartners and adult poor spellers. In D. B. Gray and J. F. Kavanagh (Eds.), *Biobehavioral measures of dyslexia* (pp. 163-176). Parkton, MD: York Press

Liberman, I. Y., & Shankweiler, D. (1979). Speech, the alphabet, and teaching to read. In L. B. Resnick & P. A. Weaver (Eds.), *Theory and practice of early reading, vol. 2* (pp. 109-134). Hillsdale, NJ: Lawrence Erlbaum Associates.

Liberman, I. Y., Shankweiler, D., Fischer, F. W., & Carter, B. (1974). Explicit syllable and phoneme segmentation in the young child. *Journal of Experimental Child Psychology, 18,* 201-212.

Liberman, I. Y., Shankweiler, D., & Liberman, A. M. (1989). The alphabetic principle and learning to read. In D. Shankweiler & I. Y. Liberman (Eds.), *Phonology and reading disability: Solving the reading puzzle.* IARLD Monograph Series, Ann Arbor, MI: University of Michigan Press.

Lundberg, I., Frost, J., & Petersen, O.-P. (1988). Effects of an extensive program for stimulating phonological awareness in preschool children. *Reading Research Quarterly, 23,* 263-284.

Mattingly, I. G. (1992). Linguistic awareness and orthographic form. In L. Katz & R. Frost (Eds.), *Orthography, phonology, morphology, and meaning* (pp. 11-26). Amsterdam: Elsevier Science Publishers.

Montessori, M. (1964). *The Montessori method.* New York: Schocken Books.

Morais, J. (1991). Constraints on the development of phonemic awareness. In S. A. Brady & D. P. Shankweiler (Eds.), *Phonological processes in literacy: A tribute to Isabelle Y. Liberman.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Morais, J., Cary, L., Alegria, J., & Bertelson, P. (1979). Does awareness of speech as a sequence of phonemes arise spontaneously? *Cognition, 7,* 323-331.

Ognjenović, V., Lukatela, G., Feldman, L. B., & Turvey, M. T. (1973). Misreadings by beginning readers of Serbo-Croatian. *Quarterly Journal of Experimental Psychology, 35A,* 97-109.

Perin, D. (1983). Phonemic segmentation and spelling. *British Journal of Psychology, 74,* 129-144.

Perfetti, C. A. (1985). *Reading ability.* New York: Oxford University Press.

Read, C. (1971). Pre-school children's knowledge of English phonology. *Harvard Educational Review, 41,* 1-34.

Read, C. (1986). *Children's creative spelling.* London: Routledge and Kegan Paul.

Richardson, E., & DiBenedetto, B. (1986). Decoding skills test. Parkton, MD: York Press.

Rohl, M., & Tunmer, W. E. (1988). Phonemic segmentation skill and spelling acquisition. *Applied Psycholinguistics, 9,* 335-350.

Rubin, H. (1988). Morphological knowledge and early writing ability. *Language and Speech, 31,* 337-355.

Shankweiler, D. (1992). Surmounting the Consonant Cluster in Beginning Reading and Writing. Presentation at AERA annual meeting, April 1992.

Shankweiler, D., Crain, S., Brady, S., & Macaruso, P. (1992). Identifying the causes of reading disability. In P. B. Gough, L. C. Ehri, & R. Treiman (Eds.), *Reading acquisition* (pp. 275-305). Hillsdale, NJ: Lawrence Erlbaum Associates.

Shankweiler, D., & Liberman, I. Y. (1972). Misreading: A search for causes. In J. F. Kavanagh & I. G. Mattingly (Eds.), *Language by ear and by eye: The relationships between speech and reading* (pp. 293-317). Cambridge, MA: MIT Press.

Treiman, R. (1985). Onsets and rimes as units of spoken syllables: Evidence from children. *Journal of Experimental Child Psychology, 39,* 161-181.

Treiman, R., & Zukowski, A. (1991). Levels of phonological awareness. In S. A. Brady & D. P. Shankweiler (Eds.), *Phonological processes in literacy: A Tribute to Isabelle Y. Liberman.* Hillsdale, NJ: Lawrence Erlbaum Associates.

BEST COPY AVAILABLE

Treiman, R. (1991). Children's spelling errors on syllable-initial consonant clusters. *Journal of Educational Psychology, 83,* 346-360.

Treiman, R. (1992). The role of intrasyllabic units in learning to read and spell. In P. B. Gough, L. C. Ehri, & R. Treiman (Eds.), *Reading acquisition* (pp. 65-106). Hillsdale, NJ: Lawrence Erlbaum Associates.

Treiman, R. (1993). *Beginning to spell: A study of first grade children.* New York: Oxford University Press.

Venezky R. L. (1970). *The structure of English orthography.* The Hague: Mouton.

Zifcak, M. (1981). Phonological awareness and reading acquisition. *Contemporary Educational Psychology, 6,* 117-126.

FOOTNOTES

*L. Katz & R. Frost (Eds.), *Orthography, phonology, morphology, and meaning* (pp. 179-192). Amsterdam: Elsevier Science Publishers (1992).

†Also University of Connecticut, Storrs.

‡University of Connecticut, Storrs.

[1]For example, it has often been noted that aspirate and inaspirate /p/, /t/ and /k/ are not distinguished in English spelling. In the word COCOA, for example, both the initial and medial consonant are spelled alike although phonetically and acoustically they are different.

[2]Often children who have had little or no formal instruction attempt to write words using the letters that they know, together with their their conceptions of the phonetic values of the letters and the segmental composition of the words they wish to write. This phenomenon has been studied extensively by Read (1986). The question of whether invented spellings can regularly be elicited from children with varied educational and family backgrounds was addressed by Zifcak (1981). In a study of 23 inner-city six year olds from blue-collar families, it was found that nearly all the children were willing to make up spellings for words though most had little knowledge of the standard orthography.

[3]These results are in full agreement in this respect with those of Treiman (1993), who carried out a comprehensive study of spelling in six year olds. The findings of both studies support the caveat that one should not be too quick to attribute children's spelling errors to the irregularities of English orthography.

[4]Spelling was correlated with reading real words, .91 and .81, respectively, based on two independent measures of reading.

[5]The onset consists of the string of consonants preceding the vowel nucleus. When the onset consists of a single consonant, as in the example of CAR, Treiman (1985) showed that children may treat it as a segment distinct from the remainder of the syllable, which corresponds to the rime. At the same time, they are unable to decompose the rime into separable components. An invented spelling, like CR for CAR or BL for BELL is consistent with such partial knowledge of the internal structure of the syllable.

[6]Adams (1990), Ehri (1989; Ehri & Wilce, 1987) and Treiman (in press) reach a similar conclusion.

Haskins Laboratories Status Report on Speech Research
1993, SR-113, 145-152

Word Superiority in Chinese*

Ignatius G. Mattingly† and Yi Xu†

In a line of research that began with Cattell (1886), it has been demonstrated that words play a special role in the recognition of text. A letter string that forms a word is recognized faster and more accurately than a nonword string; a letter is recognized faster and more accurately if it is presented as part of a word than if it is presented alone or as part of nonword (e.g., Reicher, 1969; for a review, see Henderson, 1982). This constellation of findings, often referred to as "word superiority," suggests that the reader does not simply process the text letter by letter, and that words are crucial. The letters in a word are processed so automatically that the reader is unaware of recognizing them, and when he is required to report a letter presented in a word, he finds it most efficient to infer the identity of the letter from that of the word.

Most of these experiments, however, have been carried out for writing systems like that of English, in which the "frame" units (W. S.-Y. Wang, 1981) explicitly correspond to linguistic words. However, some writing systems lack this property, notably that of Chinese. The frames of the Chinese system, the characters, correspond to monomorphemic syllables, rather than to words, even though it is true that because there are many monosyllabic words in Chinese, there are many characters that stand for words as well as for syllables. But most Chinese words are polymorphemic and are therefore written with strings of two or more characters, and these strings are not specially demarcated in text. Moreover, there are many bound morphemes in Chinese, and a character for such a morpheme occurs *only* as an element of a character string.

It is of some importance to establish whether word superiority is observable in writing systems in which the frames are not words. If word superiority is not found in these systems, we would have to view the word superiority found for word-frame systems as merely orthographic, to be attributed, perhaps, to the reader's long experience in dealing with this particular kind of frame. In the case of Chinese, we would then expect to find evidence for the superiority of morphemic syllables. On the other hand, if word-superiority *is* found even when the frames of the writing system do not correspond to words, we would have to say that the superiority of the word must depend essentially on its linguistic rather than its orthographic status.

At least two other investigators have investigated word superiority in Chinese. C.-M. Cheng (1981) compared the accuracy with which a briefly presented target character could be identified in real-word and in nonword two-character strings. (A nonword consisted of two valid characters that did not constitute a real word.) A string containing the target character was presented either preceding or following a "distractor" string, and the subject had to report whether the target character was in the first or the second string. Performance was better for words than for nonwords, and better for high-frequency words than for low-frequency words.

Cheng also carried out a second experiment, parallel to the first, in which the targets were radicals presented as components of real characters, or of pseudocharacters, or of noncharacters. Pseudocharacters were created by interchanging radicals in two real characters, keeping their position within the character constant; noncharacters were created by interchanging radicals and locating them in orthographically impossible positions. The character containing the target was presented either preceding or following a distractor character, and subjects had to report whether the target was in the first or the second character. Performance was better for real characters than for pseudocharacters, and better for pseudocharacters than for noncharacters. This result is con-

This work was supported by NICHD Grant HD-01994 to Haskins Laboratories.

sistent with a word-superiority effect for monomorphemic words; however, as Hoosain (1991) suggests, it is not conclusive. Because the real characters stood for morphemic syllables as well as for words, the result is equally consistent with morphemic-syllable superiority. It would be desirable to repeat the experiment, comparing characters standing for bound morphemes with characters for free morphemes.

I.-M. Liu (1988) asked subjects to pronounce the character occurring in first, second, or third position in word strings, in pseudoword strings or in isolation, and measured reaction time. (The position of a character in isolation apparently refers to the position of the same character when presented in a word or pseudoword.) Characters in real words were pronounced faster than characters in pseudowords in some though not in all positions. However, characters in isolation were pronounced as fast as, and in some positions faster than, characters in words. Thus, if an advantage for the real-word context over isolation is considered essential for word superiority, Chinese may not properly be said to exhibit this phenomenon and Cheng's first experiment may, as Liu argues, merely demonstrate "compound-word superiority."

But perhaps it is not reasonable to expect the real-word context to be superior to isolation if the target characters may themselves be words. We would not be surprised to find the advantage of words over single letters for English breaking down if *I* or *a*, which happen to be words as well as letters, were the targets. Analysis of Liu's results for target characters standing for bound morphemes might clarify the issue.

The experiment reported here explores further the phenomenon of word-superiority in Chinese. We asked whether a character is recognized faster when part of a two-character word than when part of a two-character pseudoword. Our experiment thus complements Cheng's (1981) first experiment, in which accuracy of identification was the dependent variable. The paradigm we used was adapted from Meyer, Schvaneveldt, and Ruddy (1974) and has already been used for Chinese by C.-M. Cheng and S.-I. Shih (1988). In each trial in our experiment, a Chinese subject saw a sequence of two characters on a monitor screen. This sequence might consist of two genuine Chinese characters, which might form either a real word or a pseudoword. In either case the subject was to respond, "Yes." Alternatively, the sequence might consist of a genuine character preceded or followed by a pseudocharacter, in which case the subject was to respond, "No." Reaction time was measured for all responses.

Methods

Design To make possible the various critical comparisons in which we were interested, a rather complex design was used; see Table I. There were three main types of two-character sequences: Real bimorphemic Chinese words; pseudowords, each, like Cheng's (1981) nonwords, consisting of two real characters that did not form a real word; and pseudocharacter sequences, each consisting of a real character preceded or followed by a pseudocharacter. The real characters were drawn from the inventory used in the People's Republic of China, and thus included some "simplified" characters not used elsewhere. The pseudocharacters, like Cheng's, each consisted of a genuine, appropriately located semantic radical and a genuine, appropriately located phonetic radical that do not actually occur together in the Chinese character inventory.

Table 1. *Experimental design.*

Sequence Type:		Real Words (128)								Pseudowords (128)				Pseudocharacter sequences (256)			
Word F:		High F (64)				Low F (64)											
Char F:		HH	HL	LH	LL	HH	HL	LH	LL	HH	HL	LH	LL	HN	LN	NH	NL
		16	16	16	16	16	16	16	16	32	32	32	32	64	64	64	64
	A	4	4	4	4	4	4	4	4	8	8	8	8	16	16	16	16
Subj.	B	4	4	4	4	4	4	4	4	8	8	8	8	16	16	16	16
Group:	C	4	4	4	4	4	4	4	4	8	8	8	8	16	16	16	16
	D	4	4	4	4	4	4	4	4	8	8	8	8	16	16	16	16

There were 128 real words, half of which were of relatively high frequency and half, of relatively low frequency. Each of these two *word*-frequency sets was divided into four secondary sets with the four possible *character*-frequency patterns: high-high, high-low, low-high, low-low. We relied on H. Wang et al. (1986) for information about word frequencies and character frequencies. Finally, each of these eight secondary sets was divided arbitrarily into four tertiary four-word sets A, B, C, D. There were in all 32 such tertiary sets. The average number of strokes per character was kept approximately equal across these tertiary sets.

There were 128 pseudowords. For each of the tertiary sets, four pseudowords were formed by swapping initial characters within the set, discarding any resulting real words. Thus, pseudowords were perfectly-matched with real words with respect to character frequency.

There were 256 pseudocharacter sequences, each consisting of a real character and a pseudocharacter. Four pseudocharacter sequences were derived by swapping the semantic radicals of the word-initial characters within each of the original tertiary sets, discarding any resulting real characters. Four more pseudocharacter sequences were formed by repeating this operation with word-final characters.

From these materials, four different but equivalent tests were compiled. Each test included 32 real words, 32 pseudowords, and 64 pseudocharacter sequences. In a particular test, the real words, the pseudowords, and the pseudocharacter sequences were composed of unrelated tertiary sets. For example, one test consisted of set A real words, set B pseudowords, pseudocharacter sequences with word-initial pseudocharacters based on set C, and pseudocharacter sequences with word-final pseudocharacters based on set B. None of the 256 characters occurred more than once within a test, and each test was balanced with respect to word frequency, character-frequency pattern within a sequence, ordinal position of pseudocharacters, and number of strokes per character. Across the four tests, each of the original 256 real characters occurred equally often in a real word and in a pseudoword, and equally often in a real character sequence and in a pseudocharacter sequence.

Subjects There were 53 subjects. All were speakers of Mandarin and graduate students or spouses of graduate students at the University of Connecticut. All had been born and educated in the People's Republic of China, and were thus familiar with the simplified characters. Subjects were paid for their participation in the experiment.

Procedure. Subjects were divided arbitrarily into four equal groups, and each group received a different test. Each subject was tested separately. A subject was told that on each trial in the test, he would see a sequence of two characters on the Macintosh computer monitor. If he was sure that both were genuine Chinese characters, he was to press the key designated as "Yes." Otherwise, he was to press the "No" key. The next trial began two seconds after the subject's response, or, if he failed to respond, two seconds after the sequence had appeared. Before the experiment began, the subject was given a 24-trial practice session with feedback.

Responses and reaction times were automatically recorded by the computer, using a program written by Leonard Katz and slightly modified by us. Reaction time for a trial was measured from the instant the two characters began to be written on the monitor screen. This measurement were subject to an error of ±8.33 msec. because the write-time could not be known with any greater accuracy.

Results

The data for 13 subjects who responded with less than 90% accuracy were excluded from further analysis. For the remaining 40 subjects, the accuracy rates were 98.5% for real words, 92.2% for pseudowords, and 92.3% for pseudocharacter sequences.

Reaction-time data for the pseudocharacter sequences, for which the correct response was "No," are shown in Figure 1. Reaction times were shorter for pseudocharacter-initial than for pseudocharacter-final sequences. They were also shorter when the real character in the sequence was of high frequency than when it was of low frequency. However, real-character frequency had a much smaller effect for pseudocharacter-initial sequences than for pseudocharacter-final ones.

An analysis of variance was carried out on the pseudocharacter sequence data for which the factors were: Subject group (A/B/C/D), pseudocharacter position (initial/final), and real-character frequency (high/low). There was no effect of subject group: $F(3,36) = .729$. The effect of pseudocharacter position was highly significant: $F(1,36) = 39.58$, $p. \leq .0001$. The effect of real-character frequency was mildly significant: $F(1,36) = 6.67$, $p. < .05$. There was a highly significant interaction between pseudocharacter position and real-character frequency: $F(1,36) = 14.40$, $p. = .0005$.

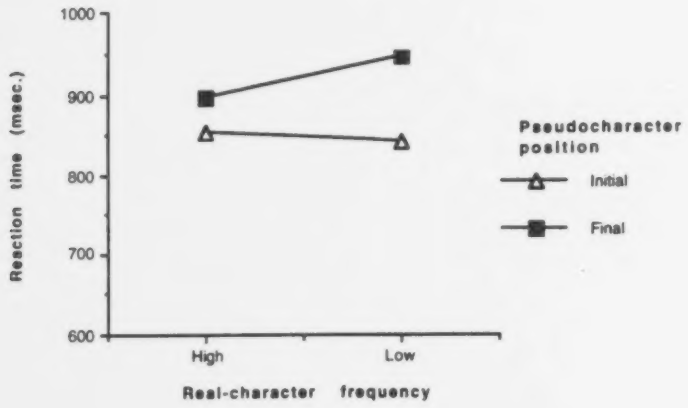


Figure 1. **Reaction times for pseudocharacter sequences.**

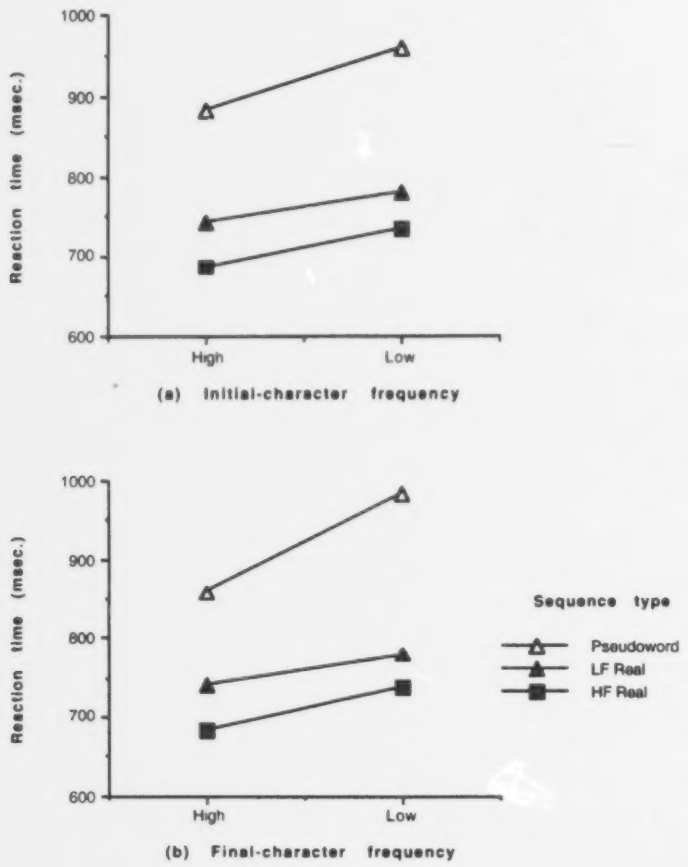


Figure 2. **Reaction times for real word and pseudoword sequences as a function of initial- and final-character frequency.**

The data for the real words and the pseudowords, for which the correct response was "Yes," are plotted in Figures 2a and 2b. Figure 2a shows the effect of varying initial character frequency; Figure 2b, the effect of varying final-character frequency. From both figures, it is apparent that reaction times are shorter for real words than for pseudowords, and shorter for high-frequency words than for low-frequency words. For both real words and pseudowords, reaction times are shorter for initial high-frequency characters than for initial low-frequency characters (Figure 2a) and similarly, shorter for final high-frequency characters than for final low-frequency characters (Figure 2b). However, reaction times for pseudowords are more affected by character frequency than are reaction times for real words.

An analysis of variance was carried out on the real and pseudoword data. The factors were: Subject group, sequence type (real word/pseudoword), initial-character frequency (high/low), and final-character frequency (high/low). There was no effect of subject group: $F(3,36) = .17$. The effect of sequence type was highly significant: $F(1,36) = 438.61$, $p. < .0001$. The effects of both character-frequency factors were highly significant: Initial, $F(1,36) = 41.47$, $p. \leq .0001$; final, $F(1,36) = 66.22$, $p. \leq .0001$. There were significant interactions between sequence type and each of the character-frequency factors: Initial: $F(1,36) = 7.29$, $p. < .05$; final: $F(1,36) = 19.31$, $p. < .0001$.

An analysis of variance was carried out on the real word data alone to determine the effects of word frequency. The factors were: Subject group, word frequency (high/low), initial-character frequency, and final-character frequency. There was no effect of subject group: $F(3,36) = .32$. The effect of word frequency was highly significant: $F(1,36) = 19.79$, $p. \leq .0001$. The effects of both character-frequency factors were highly significant: Initial, $F(1,36) = 15.03$, $p. < .0005$; final, $F(1,36) = 18.31$, $p. \leq .0001$. There was no interaction between word frequency and either of the character-frequency factors: Initial, $F(1,36) = .24$; final, $F(1,36) = .65$. There was a significant interaction among subject group, word frequency, initial character-frequency, final-character frequency: $F(3,36) = 6.56$, $p. < .005$. We believe this interaction to be artifactual, reflecting an unfortunate choice of items in one of the tertiary subsets.

The pseudoword function in Figure 2b is steeper than the pseudoword function in Figure 2a, suggesting that character frequency has more of an effect in final position than in initial position. To explore further the effect of character-frequency order, reaction time data for high-low and low-high character-frequency patterns are plotted in Figure 3. For pseudowords, reaction times for the low-high pattern are shorter than for the high-low pattern. For real words, there is no comparable effect of character-frequency pattern.

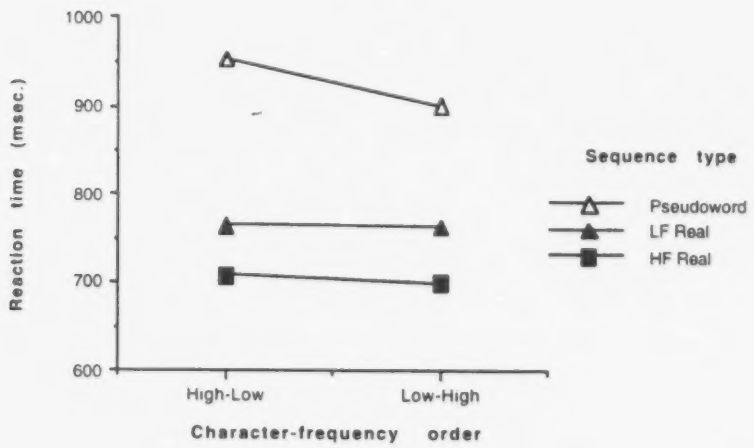


Figure 3. **Reaction times to real and pseudoword sequences as a function of character-frequency order.**

An analysis of variance was carried out on the real and pseudoword data for the high-low and low-high patterns alone. The factors were: Subject group, sequence type, and character-frequency pattern(high-low/low-high). There was no effect of subject group: $F(3,36) = .23$. The effect of sequence type was significant: $F(1,36) = 209.35$ $p. < .0001$. The effect of character-frequency pattern fell just short of significance: $F(1,36) = 3.86$, $p. = .0571$. There was a significant interaction between sequence type and character-frequency pattern: $F(1,36) = 4.39$, $p. < .05$.

An analysis of variance was also carried out for the high-low and low-high patterns in the real word data alone. The factors were; Subject group, word frequency, and character-frequency pattern. There was no effect of subject group: $F(3,36) = .24$. The effect of word frequency was significant: $F(1,36) = 21.45$, $p. \leq .0001$. There was no effect of character-frequency pattern: $F(1,36) = .15$. There was no interaction between word-frequency and character-frequency pattern: $F(1,36) = .06$. There was a significant interaction between word frequency and subject group: $F(3,36) = 6.06$, $p. < .005$. We believe this interaction has the same source as the artifactual interaction mentioned above.

Discussion

These results complement those of Cheng (1981) and provide further evidence of a word superiority effect for Chinese. The key finding is that a character is evaluated more rapidly when part of a real-word sequence than when part of a pseudoword sequence (Figure 2a,b). The advantage for real words is consistent across differences in character-frequency pattern. The magnitude of the advantage, around 200 msec., is very large.

The results also reveal something about the basis of word superiority. Let us compare the way subjects deal with sequences that are real words and sequences that are not: the pseudocharacter sequences and the pseudowords. Common to all three sequence types is the effect of character frequency (Figures 1, 2a,b). We can conclude from this simply that word superiority is not magical: Word recognition in Chinese, whatever its other properties, is mediated by character identification. The recognition of a word is evidently facilitated by previous encounters with its characters in other words. On the other hand, the fact that we found a word-frequency effect (Figure 2a,b), independent of the character-frequency effect, for a task ostensibly requiring only character recognition, suggests that word recognition cannot be simply a matter of recognizing one character at a time, then deciding that a character string is a word.

This proposal is supported by the fact that evidence of serial processing is found for pseudocharacter sequences and pseudowords, but not for words. In the case of the pseudocharacter sequences, it was found that sequences beginning with a pseudocharacter were rejected faster than those beginning with a real character, and that real-character frequency affected the latter but not the former (Figure 1). The obvious interpretation is that if the first character was genuine, the subject evaluated it, the amount of time required depending on the frequency of the character. Then he had to evaluate the pseudocharacter, which required more time, before he could reject the sequence. On the other hand, if the first character was a pseudocharacter, he could reject the sequence as soon as he could evaluate this character; there was no need to consider the second character at all. What is of interest here is simply that the subject is processing the characters serially.

As for the pseudowords, there was an effect of character-frequency order (Figure 3): The effect of character frequency was greater for the initial character than for the final character. Thus, a low-frequency character in initial position inhibited the response more than a high-frequency character in final position facilitated it; conversely, a high-frequency character in initial position facilitated the response more than a low-frequency character in final position inhibited it. We are not able to offer a conclusive explanation for this phenomenon without further experimentation, but it is plausible that when there was a low-frequency character in final position, the subject was apt to delay his evaluation, whereas he was less apt to do so when there was a low-frequency character in initial position because he had to hurry on to evaluate the final character. What is clear, however, is that the phenomenon is an order effect. It suggests that the characters in pseudowords, like those in pseudocharacter sequences, are being processed serially.

This is not the case for real words. There is no effect of character-frequency order (Figure 3). Given the order effects observed for the other two sequence types, this finding has two implications. It suggests first that the characters in real words are processed in parallel. Assume that the "logogens" (Morton, 1969) for Chinese are strings of characters corresponding to words. When a string matching a particular logogen appears in

text, all of its constituent characters will be activated at the same or nearly the same time, just like the letters of an English word (Sperling, 1969). This means that in the case of a real word, a subject in our experiment would have had only one decision to make instead of two. Having recognized the word, he would have known immediately that both characters must be genuine. Much of the advantage of real words over pseudowords may be due to this fact.

The second implication is that if logogens for two or more character strings of unequal length are activated, the logogen for the longest string is preferred. For the present experiment, this means that a subject was not free to choose whether to treat a sequence as two separate monomorphemic characters or as a single bimorphemic word. Rather, bimorphemic logogens automatically took precedence if activated. Without this "Longest String Principle," the subject would have been free to waste time by processing the real words serially, just as if they were pseudowords. This principle also explains why readers of Chinese can read rapidly despite the absence of explicit word boundary markers in the text.

Conclusion

Word superiority has been demonstrated for Chinese, and it has been argued that it depends in great part on the reader's ability to process the characters of a word in parallel. Of course, it may be that Liu (1988) is right to insist that experiments like this one and the first experiment in Cheng (1981) demonstrate merely "compound-word superiority"; the superiority of monomorphemic words remains in question. But even compound-word superiority is of great theoretical importance. Because there are no word boundaries in Chinese writing, our results are evidence that word superiority generally depends not on orthographic experience, but on linguistic experience.

REFERENCES

Cattell, J. M. (1886). The time taken up by cerebral operations. *Mind, 11,* 220-242.

Cheng, C.-M. (1981). Perception of Chinese characters. *Acta Psychologica Taiwanica,* 23, 137-153.

Cheng, C.-M., & Shih, S.-I. (1988). The nature of lexical access in Chinese: Evidence from experiments on visual and phonological priming in lexical judgment. In I.-M. Liu, H.-C. Chen, & M. J. Chen (Eds.),*Cognitive aspects of the Chinese Language, Volume 1* (pp. 1-13). Hong Kong: Asian Research Service.

Henderson, L. (1982). *Orthography and word recognition in reading.* London: Academic Press.

Hoosain, R. (1991). *Psycholinguistic implications for linguistic relativity. A case study of Chinese.* Hillsdale, NJ: Lawrence Erlbaum.

Liu, I.-M. (1988). Context effects on word/character naming: Alphabetic versus logographic languages. In I.-M. Liu, H.-C. Chen, & M. J. Chen (Eds.), *Cognitive aspects of the Chinese Language, Volume 1* (pp. 81-92). Hong Kong: Asian Research Service.

Meyer, D. E., Schvaneveldt, R. W., & Ruddy, M. (1974). Functions of graphemic and phonemic codes in visual word recognition. *Memory & Cognition, 2,* 309-321.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review, 76,* 165-178.

Reicher, G. M. (1969). Perceptual recognition as a function of the meaningfulness of the stimulus material. *Journal of Experimental Psychology, 81,* 275-280.

Sperling, G. (1969). Successive approximations to a model for short-term memory. In R. N. Haber (Ed.), *Information processing approaches to visual perception* (pp. 32-37). New York: Holt, Rinehart & Winston.

Wang, H., Chang, B., Li, Y., Lin, L., Lin, J., Sun, Y., Wang, Z., Yu, Y., Zhang, J., & Li, D. (1986). *Frequency dictionary of contemporary Chinese.* Beijing: Beijing Language Institute Press.

W. S.-Y. Wang (1981) Language structure and optimal orthography. In O. L. Tzeng & H. Singer (Eds.), *Perception of print: Reading research in experimental psychology* (pp. 223-236). Hillsdale, NJ: Lawrence Erlbaum.

FOOTNOTES

*Presented at the Sixth International Symposium on Cognitive Aspects of the Chinese Language, Taipei, Taiwan, September 1-4, 1993.

†Also Department of Linguistics, University of Connecticut, Storrs.

Haskins Laboratories Status Report on Speech Research
1993, SR-113, 153-170

Prelexical and Postlexical Strategies in Reading: Evidence from a Deep and a Shallow Orthography*

Ram Frost†

The validity of the orthographic depth hypothesis (ODH) was examined in Hebrew by employing pointed (shallow) and unpointed (deep) print. Experiments 1 and 2 revealed larger frequency effects, and larger semantic priming effects in naming, with unpointed print than with pointed print. In Experiments 3 and 4 subjects were presented with Hebrew consonantal strings that were followed by vowel marks appearing at stimulus onset asynchronies ranging from 0 ms (simultaneous presentation) to 300 ms from the onset of consonant presentation. Subjects were inclined to wait for the vowel marks to appear even though the words could be named unequivocally using lexical phonology. These results suggest that prelexical phonology is the default strategy for readers in shallow orthographies, providing strong support for the ODH.

Most early studies of visual word recognition were carried out in the English language. This state of affairs was partly due to an underlying belief that reading processes (as well as other cognitive processes) are universal, and therefore studies in English are sufficient to provide a complete account of the processes involved in recognizing printed words. In the last decade, however, studies in orthographies other than English have become more and more prevalent. These studies have in common the view that reading processes cannot be explained without considering the reader's linguistic environment in general, and the characteristics of his writing system in particular.

Various writing system have evolved over time in different cultures. These writing systems, whether logographic, syllabic, or alphabetic, typically reflect the language's unique phonology and morphology, representing them in an effective way (Mattingly, 1992; Scheerer, 1986). The match between writing system and language insures a degree of efficiency for the reading and the writing process (for a discussion, see Katz & Frost, 1992). Theories of visual recognition differ in their account of how the different characteristics of writing systems affect reading performance. One such characteristic that has been widely investigated is the way the orthography represents the language's surface phonology.

The transparency of the relation between spelling and phonology varies widely between orthographies. This variance can often be attributed to morphological factors. In some languages (e.g., in English), morphological variations are captured by phonologic variations. The orthography, however, was designed to preserve primarily morphologic information. Consequently, in many cases, similar spellings denote the same morpheme but different phonologic forms: The same letter can represent different phonemes when it is in different contexts, and the same phoneme can be represented by different letters. The words "steal" and "stealth," for example, are similarly spelled because they are morphologically related. Since in this case, however, a morphologic derivation resulted in a phonologic variation, the cluster "ea" represents both the sounds [i] and [E]. Thus, alphabetic orthographies can be classified

This work was supported in part by a grant awarded to the author by the Basic Research Foundation administered by the Israel Academy of Science and Humanities, and in part by the National Institute of Child Health and Human Development Grant HD-01994 to Haskins Laboratories. I am indebted to Len Katz for many of the ideas that were put forward in this article, to Orna Moshel and Itchak Mendelbaum for their help in conducting the experiments, and to Ken Pugh, Bruno Repp, Charles Perfetti, Albrecht Inhoff, and Sandy Pollatsek for their comments on earlier drafts of this paper.

according to the transparency of their letter to phonology correspondence. This factor is usually referred to as "orthographic depth" (Katz & Feldman, 1981; Klima, 1972; Liberman, Liberman, Mattingly, & Shankweiler, 1980; Lukatela, Popadić, Ognjenović, & Turvey, 1980). An orthography that represents its phonology unequivocally following grapheme-phoneme simple correspondences is considered shallow, while in a deep orthography the relation of orthography to phonology is more opaque.

The effect of orthographic depth on reading strategies has been the focus of recent and current controversies (e.g., Besner & Smith, 1992; Frost, Katz, & Bentin, 1987). In general the argument revolves around the question of whether differences in orthographic depth lead to differences in processing printed words. What is called the orthographic depth hypothesis (ODH) suggests that it does. The ODH suggests that shallow orthographies can easily support a word recognition process that involves the printed word's phonology. This is because the phonologic structure of the printed word can be easily recovered from the print by applying a simple process of grapheme-to-phoneme conversion (GPC). For example, in the Serbo-Croatian writing system the letter-to-phoneme correspondence is consistent and the spoken language itself is not phonologically complex. The correspondence between spelling and pronunciation is so simple and direct that a reader of this orthography can expect that phonological recoding will always result in an accurate representation of the word intended by the writer. A considerable amount of evidence now supports the claim that phonological recoding is extensively used by readers of Serbo-Croatian (see, Carello, Turvey, & Lukatela, 1992, for a review). In contrast to shallow orthographies, deep orthographies like English or Hebrew encourage readers to process printed words by referring to their morphology via the printed word's visual-orthographic structure. In deep orthographies lexical access is based mainly on the word's orthographic structure, and the word's phonology is retrieved from the mental lexicon. This is because the relation between the printed word and its phonology are more opaque and prelexical phonologic information cannot be easily generated (e.g., Frost et al. 1987; Frost & Bentin, 1992b; Katz & Feldman, 1983).

The ODH's specific predictions mainly refer to the way a printed word's phonology is generated in the reading process. Because readers of shallow orthographies have simple, consistent, and relatively complete connections between graphemes and sub-word pronunciation, they can recover most of a word's phonological structure prelexically, by assembling it directly from the printed letters. In contrast, the opaque relation of subwords segments and phonemes in deep orthographies prevents readers from using prelexical conversion rules. For these readers, the more efficient process of generating the word's phonologic structure is to rely on a fast visual access of the lexicon and to retrieve the word's phonology from it. Thus, phonology in this case is lexically addressed, not prelexically assembled. Thus, according to the ODH, the difference between deep and shallow orthographies is the amount of lexical involvement in pronunciation. This does not necessarily entail specific predictions concerning lexical decisions and lexical access, or how meaning is accessed from print.

The specific predictions of the ODH must be discussed with reference to the tools of investigation employed by the experimenter. The issue to be clarified here is what serves as a valid demonstration that phonology is mainly prelexical (assembled) or postlexical (addressed). In general, measures of latencies and error rates for lexical decisions or naming are monitored. The idea is that lexical involvement in pronunciation leaves characteristic traces. The first question to be examined is, therefore, whether the lexical status of a word affects naming latencies. Lexical search results in frequency effects and in lexicality effects: Frequent words are named faster than nonfrequent words, and words are named faster than nonwords (but see Balota & Chumbley, 1984, for a discussion of this point). Thus, one trace of postlexical phonology is that naming latencies and lexical decision latencies are similarly effected by the lexical status of the printed stimulus (e.g., Katz & Feldman, 1983). If phonology is assembled from print prelexically, smaller effects related to the word's lexical status should be expected, and consequently lexical status should affect naming and lexical decisions differently. A second method of investigation involves the monitoring of semantic priming effects in naming (see Lupker, 1984; Neely, 1991, for a review). If pronunciation involves postlexical phonology, strong semantic priming effects will be revealed in naming. In contrast, if pronunciation is carried out using mainly prelexical phonology, naming of target words would be less facilitated by semantically related primes.

Keeping the above tools of investigation in mind, the exact predictions of the ODH can be now fully

described. Two versions of the hypothesis exist in the current literature. What can be called the *strong* ODH claims that in shallow orthographies, the complete phonological representations can be derived exclusively from only the prelexical translation of subword spelling units (letter or letter clusters) into phonological units (phonemes, phoneme clusters, and syllables). According to this view, readers of shallow orthographies perform a phonological analysis of the word based only on a knowledge of these correspondences. Rapid naming, then, is a result of this analytic process *only*, and does not involve any lexical information (see Katz & Frost, 1992, for a discussion). In contrast to the strong ODH, a weaker version of the ODH can be proposed. According to the weak version of the ODH, the phonology needed for the pronunciation of printed words comes *both* from prelexical letter-phonology correspondences and from stored lexical phonology. The latter is the result of either an orthographic addressing of the lexicon (i.e., from a whole-word or whole-morpheme spelling pattern to its stored phonology), or from a partial phonologic representation that was assembled from the print and was unequivocal enough to allow lexical access. The degree to which the prelexical process is active is a function of the depth of the orthography; prelexical analytic processes are more functional in shallow orthographies. Whether or not prelexical processes actually dominate orthographic processing for any particular orthography is a question of the demands the two processes make on the reader's processing resources (Katz & Frost, 1992).

It is easy to show that the strong form of the ODH is untenable. It is patently insufficient to account for pronunciation even in shallow orthographies like Spanish, Italian, or Serbo-Croatian. This is so because these orthographies do not represent syllable stress and, even though stress is often predictable, this is not always the case. For example, in Serbo-Croatian, stress for two-syllable words always occurs on the first syllable, but not always for words of more than two syllables. These words can be pronounced correctly only by reference to lexically stored information. The issue of stress assignment is even more problematic in Italian, where stress patterns are much less predictable. In Italian, many pairs of words differ only in stress which provides the intended semantic meaning (Colombo & Tabossi, 1992; Laudanna & Caramazza, 1992).

Several studies have argued for the obligatory involvement of prelexical phonology in Serbo-Croatian (e.g., Turvey, Feldman, & Lukatela, 1984) or in English (Perfetti, Bell, & Delaney, 1988; Perfetti, Zhang, & Berent, 1992; Van Orden, 1987). Note than the weaker version of the ODH is not inconsistent with these claims as long as it is not claimed that naming is achieved *exclusively* by prelexical analysis. All alphabetic orthographies may make some use of prelexically derived phonology for word recognition. According to the weak form of the ODH, shallow orthographies should make more use of it than deep orthographies, because prelexical phonology is more readily available in these orthographies. Substantial prelexical phonology may be generated inevitably when reading Serbo-Croatian, for example. This phonological representation may be sufficient for lexical access, but not necessarily for pronunciation. The complete analysis of the word's phonologic and phonetic structure may involve lexical information as well (Carello et al., 1992).

Evidence concerning the validity of the ODH comes from within- and cross-language studies. But note that single-language experiments are adequate only for testing the strongest form of the ODH. The strong ODH requires lexical access and pronunciation to be accomplished entirely on the basis of phonological information derived from correspondences between subword spelling and phonology. That is, the reader's phonological analysis of the printed word, based on his or her knowledge of subword letter-sound relationships, is the only kind of information that is allowed. Thus, merely showing lexical involvement in pronunciation in a shallow orthography would provide a valid falsification of the strong ODH (Seidenberg & Vidanović, 1985; Sebastián-Gallés, 1991). However, the demonstration of lexical effects in shallow orthographies cannot in itself be considered evidence against the weak ODH, and a different methodology is necessary to test it. Because it refers to the *relative* use of prelexical phonologic information in word recognition in different orthographies, experimental evidence bearing on the weak ODH should come mainly from cross-language studies. More important, because the weak ODH claims that readers in shallow orthographies do *not* use the assembled routine exclusively, the evidence against it cannot come from single-language studies by merely showing the use of the addressed routine in shallow orthographies (e.g., Sebastián-Gallés, 1991).

The experimental evidence supporting the weak ODH is abundant. Katz and Feldman (1983) compared semantic priming effects in naming and lexical decision in English and Serbo-Croatian, and demonstrated that while semantic facilitation

was obtained in English for both lexical decision and naming, in Serbo-Croatian semantic priming facilitated only lexical decision. Similarly, a comparison of semantic priming effects in naming in English and Italian showed greater effects in the deeper English than in the shallower Italian orthography (Tabossi & Laghi, 1992). A study by Frost et al. (1987) involved a simultaneous comparison of three languages, Hebrew, English, and Serbo-Croatian, and confirmed the hypothesis that the use of prelexical phonology in naming varies as a function of orthographic depth. Frost et al. showed that the lexical status of the stimulus (its being a high- or a low-frequency word or a nonword) affected naming latencies in Hebrew more than in English, and in English more than in Serbo-Croatian. In a second experiment, Frost et al. showed a relatively strong effect of semantic facilitation in Hebrew (21 ms), a smaller but significant effect in English (16 ms), and no facilitation (0 ms) in Serbo-Croatian.

Frost and Katz (1989) examined the effects of visual and auditory degradation on the ability of subjects to match printed to spoken words in English and Serbo-Croatian. They showed that both visual and auditory degradation had a much stronger effect in English than in Serbo-Croatian, regardless of word frequency. These results were explained by an extension of an interactive model which rationalized the relationship between the orthographic and phonologic systems in terms of lateral connections between the systems at all of their levels. The structure of these lateral connections was determined by the relationship between spelling and phonology in the language: simple isomorphic connections between graphemes and phonemes in the shallower Serbo-Croatian, but more complex, many-to-one, connections in the deeper English. Frost and Katz argued that the simple isomorphic connections between the orthographic and the phonologic systems in the shallower orthography enabled subjects to restore both the degraded phonemes from the print and the degraded graphemes from the phonemic information, with ease. In contrast, in the deeper orthography, because the degraded information in one system was usually consistent with several alternatives in the other system, the buildup of sufficient information for a unique solution to the matching judgment was delayed, so the matching between print and degraded speech, or between speech and degraded print, was slowed.

The psychological reality of orthographic depth is not unanimously accepted. Although it is generally agreed that the relation between spelling and phonology in different orthographies might affect reading processes (especially reading acquisition) to a certain extent, there is disagreement as to the relative importance of this factor. What I will call here "the alternative view" argues that the primary factor determining whether or not the word's phonology is assembled pre- or post-lexically is not orthographic depth but word frequency. The alternative view suggests that in *any* orthography, frequent words are very familiar as visual patterns. Therefore, these words can easily be recognized through a fast visually-based lexical access which occurs before a phonologic representation has time to be generated prelexically from the print. For these words, phonologic information is eventually obtained, but only postlexically, from memory storage. According to this view, the relation of spelling to phonology should not affect recognition of frequent words. Since the orthographic structure is not converted into a phonologic structure by use of grapheme-to-phoneme conversion rules, the depth of the orthography does not play a role in the processing of these words. Orthographic depth exerts some influence, but only on the processing of low-frequency words and nonwords. Since such verbal stimuli are less familiar, their visual lexical access is slower, and their phonology has enough time to be generated prelexically (Baluch & Besner, 1991; Tabossi & Laghi, 1992; Seidenberg, 1985).

A few studies involving cross-language research support the alternative view. Seidenberg (1985) demonstrated that in both English and Chinese, naming frequent printed words was not effected by phonologic regularity. This outcome was interpreted to mean that, in logographic as in alphabetic orthographies, the phonology of frequent words was derived postlexically, after the word had been recognized on a visual basis. Moreover, in another study, Seidenberg and Vidanović (1985) found similar semantic priming effects in naming frequent words in English and Serbo-Croatian, suggesting again that the addressed routine plays a major role even in the shallow Serbo-Croatian.

Several studies in Japanese were interpreted as providing support for the alternative view. Although these studies were carried out within one language, they could in principle furnish evidence in favor or against the weak ODH because they examined reading performance in two different writing systems that are commonly used in Japanese- the deep logographic Kanji and the shallower syllabic Hiragana and Katakana. However, the relevance of these studies to the debate concerning the weak ODH is questionable, as

I will point out below. Besner and Hildebrant (1987) showed that words that were normally written in Katakana were named faster than words written in Katakana that were transcribed from Kanji. They argued that these results suggest that readers of the shallow Katakana did not name the transcribed words using the assembled routine, as otherwise no difference should have emerged in naming the two types of stimuli. In another experiment, Besner, Patterson, Lee, and Hildebrant (1992) showed that naming words that are normally seen in Katakana were named slower if they were written in Hiragana and vice versa. This outcome suggests that orthographic familiarity plays a role even in reading the shallower syllabic Japanese orthographies. Taken together, the results from Japanese provide strong evidence against the strong ODH, but they do not contradict the weak ODH. Although these studies compare reading performance in two writing systems, they merely show that readers of the shallower Japanese syllabary do not use the assembled routine exclusively, but use the addressed routine as well. These conclusions, however, are actually the basic tenets of the weak ODH.

More damaging to the weak ODH is a recent study by Baluch and Besner (1991), who employed a within-language between orthography design in Persian, similar to the one in Japanese. In this study the authors took advantage of the fact that some words in Persian are phonologically transparent whereas other words are phonologically opaque. This is because three of the six vowels of written Persian are represented in print as diacritics and three are represented as letters. Because (as in Hebrew) fluent readers do not use the pointed script, words that contain vowels represented by letters are phonologically transparent, whereas words that contain vowels represented by diacritics are phonologically opaque. Baluch and Besner demonstrated similar semantic priming effects in naming phonologically transparent and phonologically opaque words of Persian, provided that nonwords were omitted from the stimulus list. These results were interpreted to suggest that the addressed routine was used in naming both types of words.

Thus, when the weak ODH and the traditional alternative view are contrasted, it appears that they differ in one major claim concerning the preferred route of the cognitive system for obtaining the printed word's phonology. The alternative view suggests that in general *visual* access of the lexicon is "direct" and requires fewer cognitive resources. Moreover, the extraction of phonological information from the mental lexicon is more or less effortless relative to the extraction of phonological information prelexically. Hence, the default of the cognitive system in reading is the use of addressed phonology. The weak ODH denies that the extraction of phonological information from the mental lexicon is effortless. On the contrary, based on findings showing extensive phonologic recoding in shallow orthographies, its working hypothesis is that if the reader can successfully employ prelexical phonological information, then it will be used first; the easier it is, the more often it will be used. Thus, the "default" of the cognitive system in word recognition is the use of prelexical rather than addressed phonology. If the reader's orthography is a shallow one with uncomplicated, direct, and consistent correspondences between letters and phonology, then the reader will be able to use such information with minimal resources for lexical access. The logic for the ODH lies in a simple argument. The so called "fast," "effortless," visual route is available for readers in *all* orthographies, including the shallow ones. If in spite of that it can be demonstrated that readers of shallow orthographies do prefer prelexical phonological mediation over visual access, it is because it is more efficient and faster for these readers.

The present study addresses this controversy by looking at a deeper orthography, namely Hebrew. In Hebrew, letters represent mostly consonants, while vowels can optionally be superimposed on the consonants as diacritical marks. The diacritical marks, however, are omitted from most reading material, and are found only in poetry, children's literature, and religious texts. In addition, like other Semitic languages, Hebrew is based on word families derived from triconsonantal roots. Therefore, many words (whether morphologically related or not) share a similar or identical letter configuration. If the vowel marks are absent, a single printed consonantal string usually represents several different spoken words. Thus, in its unpointed form, the Hebrew orthography does not convey to the reader the full phonemic structure of the printed word, and the reader is often faced with phonological ambiguity. In contrast to the unpointed orthography, the pointed orthography is a very shallow writing system. The vowel marks convey the missing phonemic information, making the printed word phonemically unequivocal. Although the diacritical marks carry mainly vowel information, they also differentiate in some instances between fricative and stop variants of consonants. Thus the presentation of vowels consid-

erably reduces several aspects of phonemic ambiguity (see Frost & Bentin, 1992b, for a discussion).

Several studies have shown that lexical decisions to Hebrew phonologically ambiguous words are given prior to any phonological disambiguation suggesting that lexical access in unpointed Hebrew is accomplished more often via orthographic representations than via phonological recodings of the orthography (Bentin & Frost, 1987; Bentin, Bargai, & Katz, 1984; Frost & Bentin, 1992a; Frost, 1992; and see Frost & Bentin, 1992b, for a review). The purpose of the present study was to show that even the reader of Hebrew who is exposed almost exclusively to unpointed print and generally uses the lexical addressed routine, prefers to use the prelexical assembled routine when possible. Note that the probability of extending the cross-language findings (e.g., Frost et al., 1987) to a within-Hebrew experimental design suffers from the fact that readers in general are used to regularly employing strategies of reading that arise from the characteristics of their orthography. Obviously, if it can be shown that even in spite of this factor Hebrew readers are willing to alter their reading strategy and adopt a prelexical routine, it will certainly provide very significant evidence in favor of the ODH.

EXPERIMENT 1

In Experiment 1 lexical decision and naming performance with pointed and unpointed print were compared. The aim of the experiment was to assess the effects of lexical factors on naming performance relative to lexical decision in the deep and shallow forms of the Hebrew orthography. This experimental design is very similar to that employed by Frost et al. (1987) in the first experiment of their multilingual study. Frost et al. found that in unpointed Hebrew the lexical status of the stimuli affected lexical decision latencies in the same way that it affected naming latencies. In contrast, in English and in Serbo-Croatian the pronunciation task was much less affected by the lexical status of the stimuli. The shallower the orthography, the more naming performance deviated from lexical decision performance. This outcome confirmed that, in unpointed Hebrew, phonology was generated mainly using the addressed routine, whereas in the shallower English orthography the assembled routine had a more significant role. Finally, in the shallowest orthography, Serbo-Croatian, the assembled routine dominated the addressed routine, resulting in a nonsignificant frequency effect. The purpose of Experiment 1 was to investigate whether a similar gradual deviation of naming from lexical decision performance would be observed in pointed relative to unpointed Hebrew. If the weak ODH is correct, then the reader of Hebrew should use the assembled routine to a greater extent when naming pointed print than when naming unpointed print.

Method

Subjects. One hundred and sixty undergraduate students from the Hebrew University, all native speakers of Hebrew, participated in the experiment for course credit or for payment.

Stimuli and Design. The stimuli were 40 high-frequency words, 40 low-frequency words, and 80 nonwords. All stimuli were three to five letters long, and contained two syllables with four to six phonemes. The average number of letters and phonemes were similar for the three types of stimuli. All words could be pronounced as only one meaningful word. Nonwords were created by altering randomly one or two letters of high- or low-frequency real words that were not employed in the experiment. The nonwords were all pronounceable and did not violate the phonotactic rules of Hebrew. In the absence of a reliable frequency count in Hebrew, the subjective frequency of each word was estimated using the following procedure: A list of 250 words was presented to 50 undergraduate students, who rated the frequency of each word on a 7-point scale from very infrequent (1) to very frequent (7). The rated frequencies were averaged across all 50 judges, and all words in the present study were selected from this pool. The average frequency of the high-frequency words was 4.0, whereas the average frequency of the low-frequency words was 2.0. Examples of pointed and unpointed Hebrew words are presented in Figure 1.

	Unpointed	Pointed
Hebrew print	פסנתר	פְּסַנְתֵּר
Phonologic structure	/ psanter /	
Semantic meaning	Piano	

Figure 1. **Example of the unpointed and the pointed forms of a Hebrew printed word.**

There were four experimental conditions: the stimuli could be presented pointed or unpointed, for naming or for lexical decision. Forty different subjects were tested in each experimental

condition. This blocked design was identical to that of the Frost et al. (1987) study.

Procedure and apparatus. The stimuli were presented on a Macintosh SE computer screen, and a bold Hebrew font, size 24, was used. Subjects were tested individually in a dimly lighted room. They sat 70 cm from the screen so that the stimuli subtended a horizontal visual angle of 4 degrees on the average. The subjects communicated lexical decisions by pressing a "yes" or a "no" key. The dominant hand was always used for the "yes" response. In the naming task, response latencies were monitored by a Mura-DX 118 microphone connected to a voice key. Each experiment started with 16 practice trials, which were followed by the 160 experimental trials presented in one block. The trials were presented at a 2.5 sec intertrial interval.

Results

Means and standard deviations of RTs for correct responses were calculated for each subject in each of the four experimental conditions. Within each subject/condition combination, RTs that were outside a range of 2 SDs from the respective mean were excluded, and the mean was recalculated. Outliers accounted for less than 5% of all responses. This procedure was repeated in all the experiments of the present study.

Mean RTs and error rates for high-frequency words, low-frequency words and nonwords in the different experimental conditions are presented in Table 1.[1] Point presentation had very little effect in the lexical decision task. In contrast, in the naming task the effect of frequency and lexical status of the stimulus were more pronounced in the unpointed presentation than in the pointed presentation. Moreover, naming was more similar to lexical decision performance in unpointed than in pointed print.

The statistical significance of these differences was assessed by an analysis of variance (ANOVA) across subjects (F1) and across stimuli (F2), with the main factors of stimulus type (high-frequency words, low-frequency words, nonwords), point presentation (pointed, unpointed), and task (naming, lexical decision). The main effect of stimulus type was significant ($F1(2,312) = 267.0$, $MS_e = 1550$, $p<0.001$, $F2(2,157) = 119.0$, $MS_e = 4705$, $p<0.001$). The main effects of point presentation and task were not significant in the subject analysis ($F1(1,156) = 1.7$, $MS_e = 21,506$, $p<0.2$; $F1(1,156) = 2.1$, $MS_e = 21,506$, $p<0.2$, respectively), but were significant in the stimulus analysis ($F2(1,157) = 63.7$, $MS_e = 3258$, $p<0.001$; $F2(1,157) = 23.1$, $MS_e = 3258$, $p<0.001$, respectively). Point presentation interacted with task ($F1(1,156) = 4.3$, $MS_e = 21,506$, $p<0.04$; $F2(1,157) = 97.9$, $MS_e = 1877$, $p<0.001$). Stimulus type interacted both with point presentation and with task ($F1(2,312) = 3.0$, $MS_e = 1550$, $p<0.05$; $F2(2,157) = 11.9$, $MS_e = 832$, $p<0.001$; $F1(2,312) = 30.0$, $MS_e = 1550$, $p<0.001$; $F2(2,157)=17.4$, $MS_e = 3258$, $p<0.001$, respectively). The three-way interaction did not reach significance. This was probably due to a similar slowing of nonword latencies relatively to the low-frequency words in the two prints. The difference between naming pointed and unpointed presentation was most conspicuous while examining the effect of word frequency. A Tukey-A post-hoc analysis ($p<0.05$) revealed that while there was a significant frequency effect in naming unpointed print (28 ms), there was no significant frequency effect in naming pointed print (9 ms). Finally, the correlation of RTs in the naming and the lexical decision task was calculated for both types of print presentation. RTs in the two tasks were highly correlated in the unpointed presentation ($r=0.56$), but much less so in the pointed presentation ($r=0.28$). This suggests that the lexical status of the stimuli affected lexical decision and naming similarly in unpointed print but not in pointed print.

Table 1. *RTs and percent errors in the lexical decision and naming tasks the for high-frequency words, low-frequency words, and nonwords in unpointed and pointed print.*

	UNPOINTED PRINT			POINTED PRINT		
	High-freq.	Low-freq.	Nonwords	High-freq.	Low-freq.	Nonwords
LEXICAL DECISION	529	617	659	545	627	664
	4%	2%	5%	5%	5%	5%
NAMING	569	597	664	541	550	604
	0%	6%	2%	0%	2%	9%

Discussion

The results of Experiment 1 replicate to a certain extent the findings of Frost et al. (1987), with a within-language, between-orthography design. The pattern of naming and lexical decision latencies in unpointed Hebrew was almost identical to the pattern obtained in the Frost et al. study. However, the relations of lexical decision to naming latencies in pointed Hebrew were similar to those obtained by Frost et al. in the shallower orthographies of English and Serbo-Croatian. The patterns of response times in naming and lexical decisions were fairy similar in unpointed print. This suggests a strong reliance on the addressed routine in naming when the vowel marks were not present. In contrast, naming deviated from the pattern obtained in the lexical decision task when the vowel marks were presented. The frequency effect almost disappeared, and the overall difference between naming high-frequency words and nonwords was considerably reduced. This outcome suggests that the presentation of vowel marks encouraged the readers to adopt a strategy of assembling the printed word phonology by using prelexical conversion rules. Note, however, that although the overall difference between high-frequency words and nonwords was smaller in pointed print than in unpointed print, the difference between low-frequency words and nonwords were fairly similar in the two conditions (54 in pointed print vs. 67 in unpointed print). This outcome is not fully consistent with a prelexical computation of printed stimuli in the shallow pointed print. A possible explanation for the unexpected slow naming latencies for pointed nonwords is that although phonology could be easily computed prelexically in pointed print, subjects were also sensitive to the familiarity of the printed stimuli. Readers in Hebrew read mostly unpointed print but have a large experience in reading pointed print as well. Because they are obviously more familiar with reading words than nonwords, the slower latencies for nonwords could be attributed to a response factors rather than factor related to the computation of phonology.

EXPERIMENT 2

The aim of Experiment 2 was to examine the effects of semantic facilitation in the naming task when pointed and unpointed words are presented. Semantic priming effects in the naming task have been used in several studies to monitor the extent of lexical involvement in pronunciation (e.g., Baluch & Besner, 1991; Frost et al. 1987; Tabossi & Laghi, 1992). In general, it has been shown that in languages with deep orthographies such as English, semantic facilitation in naming can easily be obtained, although the effects are usually smaller than the effect obtained in the lexical decision task (Lupker, 1984; Neely, 1991). In contrast, in shallow orthographies such as Serbo-Croatian, semantic priming effects were not obtained in some studies (e.g., Katz & Feldman, 1983; Frost et al., 1987), but have been shown in other studies (e.g., Carello, Lukatela, & Turvey, 1988; Lukatela, Feldman, Turvey, Carello, & Katz, 1989).

An examination of the weak version of the ODH entails, however, a comparison of semantic priming effects in a deep and a shallow orthography. Katz and Feldman (1983) showed greater semantic facilitation in English than in Serbo-Croatian. Similarly, Frost et al. (1987) demonstrated a gradual decrease in semantic facilitation from Hebrew (deepest orthography) to English (shallower) to Serbo-Croatian (shallowest). Similar results were suggested by Tabossi and Laghi (1992) who compared English to Italian. Recently, Baluch, and Besner (1991) have challenged these findings, suggesting that all those cross-language differences were obtained because nonwords were included in the stimulus lists. According to their proposal the inclusion of nonwords encouraged the use of a prelexical naming strategy in the shallower orthographies. Indeed, when nonwords were not included in the stimulus lists, no differences in semantic facilitation were found in naming phonologically opaque and phonologically transparent words in Persian.

Experiment 2 was designed to examine the weak version of the ODH using a semantic priming paradigm, while considering the hypothesis that the effects of orthographic depth are caused by the mere inclusion of nonwords in the stimuli lists. For this purpose semantic facilitation in naming target words was examined in pointed and unpointed Hebrew orthography, when only words were employed.

Methods

Subjects. Ninety-six undergraduate students from the Hebrew University, all native speakers of Hebrew, participated in the experiment for course credit or for payment.

Stimuli and design. The stimuli were 48 target words that were paired with semantically related and semantically unrelated primes. Related targets and primes were two instances of a semantic category. In order to avoid repetition

effects, two lists of stimuli were constructed. Targets presented with semantically related primes in one list were unrelated in the other list, and vice versa. Each subject was tested in one list only. There were two experimental conditions: in one condition all stimuli were pointed, and in the other they were unpointed. Forty-eight subject were tested in each condition, 24 on each list. The targets were three to five letter words and had two or three syllables. Both primes and targets were unambiguous, and each could be read as a meaningful word in only one way.

Procedure and apparatus. An experimental session consisted of 16 practice trials followed by the 48 test trials. Each trial consisted of a presentation of the prime for 750 ms followed by the presentation of the target. Subject were instructed to read the primes silently and to name the targets aloud. The exposure of the target was terminated by the subject's vocal response, and the intertrial interval was 2 seconds. The apparatus was identical to the one used in the naming task of Experiment 1.

Results

RTs in the different experimental conditions are presented in Table 2. Naming of related targets was faster than naming of unrelated targets in both pointed and unpointed print. However, semantic facilitation was twice as large with unpointed print than with pointed print.

Table 2. *Reaction times and percent errors to related and unrelated targets with pointed and unpointed print.*

	Unpointed Print	Pointed Print
Unrelated	531 (1%)	512 (2%)
Related	509 (4%)	503 (1%)
Priming Effect	22	9

The statistical significance of these effects was assessed by an ANOVA across subjects (F1) and across stimuli (F2), with the main factors of semantic relatedness (related, unrelated), and print type (pointed, unpointed). The effect of semantic relatedness was significant ($F1(1,47) = 20.9$, $MS_e = 656$, $p < 0.001$; $F2(1,94) = 33.2$, $MS_e = 363$, $p<0.001$). The effect of print type was significant in the subject analysis ($F1(1,47) = 30.8$, $MS_e = 232$, $p<0.001$), but not in the stimulus analysis ($F2(1,94) = 1.0$). More important to our hypothesis, the two-way interaction was significant in both analyses ($F1(1,47) = 6.6$, $MS_e = 207$, $p< 0.01$; $F2(1,94) = 4.7$, $MS_e = 363$, $p<0.03$). A Tukey-A post-hoc analysis of the interaction ($p<0.05$) revealed that the semantic facilitation was significant only with unpointed print, but not with pointed print.

DISCUSSION

Similar to experiment 1, naming in pointed print was found to be faster than naming in unpointed print, as revealed by the difference in RTs in the unrelated condition. However, whereas semantic relatedness accelerated naming in the related condition in unpointed print, it had a much smaller effect on accelerating naming latencies in pointed print. Thus the results of Experiment 2 suggest that semantic facilitation is stronger in the deeper than in the shallower Hebrew orthography. This outcome is in complete agreement with the findings of Frost et al. (1987), and Tabossi and Laghi (1992). Note, however, that the greater effects of semantic facilitation in pointed print than in unpointed print were obtained even though nonwords were not included in the stimulus set. These results conflict with the findings of Baluch and Besner (1991). We will refer to this in the general discussion.

EXPERIMENT 3

The major claim of the weak ODH is that in all orthographies both prelexical and lexical phonology are involved in naming. Thus, the use of lexical or prelexical phonology is not an all-or-none process but a quantitative continuum. The degree to which a prelexical process of assembling phonology predominates over the lexical routine depends on the costs involved in assembling phonology directly from the print. The ODH suggests that unless the costs are too high in terms of processing resources (as is usually the case in deep orthographies), the default strategy of the cognitive system is to assemble a phonologic code for lexical access, not to retrieve it from the lexicon following visual access. The following two experiments examined this aspect of the hypothesis by manipulating the processing costs for generating a prelexical phonological representation.

The manipulation of cost consisted of delaying the presentation of the vowel marks relative to the presentation of the consonant letters. In Experiment 3 subjects were presented with

unambiguous letter strings that could be read as one meaningful word only (i.e., only one vowel combination created a meaningful word). The letters were followed by the vowel marks, which were superimposed on the consonants, but at different time intervals ranging from 0 ms (in fact, a regular pointed presentation) to 300 ms from the onset of consonant presentation. Subjects were instructed to make lexical decisions or to name the words as soon as possible.

The vowel marks allow the easy prelexical assembly of the word's phonology using spelling-to-phonology conversion rules. However, since the letter strings were unambiguous and could be read as a meaningful word in only one way, they could be easily named using the addressed routine, by accessing the lexicon visually. In fact, as we have shown in numerous studies, this naming strategy is characteristic of reading unpointed Hebrew (see Frost & Bentin, 1992b, for a review). The question was, therefore, whether subjects are inclined to delay their response and wait for the vowel marks that are not indispensable for either correct lexical decisions or for unequivocal pronunciation. If they are willing to wait for the vowel marks to appear then it must be because they prefer the option of prelexical assembly of phonology, just as the ODH would predict. The relative use of the assembled and addressed routines was also verified by using both high-frequency and low-frequency words in the stimulus lists. It was assumed that the more subjects rely on the vowel marks for naming using GPC conversion rules, the smaller would be the frequency effect in this task. Nonwords were introduced as well to allow a baseline assessment of the lagging costs. Note that nonwords cannot be unequivocally pronounced before the vowel marks are presented. In order to correctly pronounce them, subjects have to wait the full lag period. However, since at least some of the articulation program can be launched immediately after the consonants are presented, the absolute difference in RTs between the presentation at lag 0 and at lag 300 of nonwords would reflect the actual cost in response time for lagging the vowel marks by 300 ms. Any lag effect smaller than this difference would suggest that subject did not wait the full lag period but combined both prelexical and lexical routines to formulate their response.

Method

Subjects. Ninety-six undergraduate students from the Hebrew University, all native speakers of Hebrew, participated in the experiments for course credit or for payment. Forty-eight participated in the lexical decision task and 48 in the naming task.

Stimuli and Design. The stimuli were 40 high-frequency words (mean frequency 5.0 on the previously described 1-7 scale), 40 low-frequency words (mean frequency 3.6), and 80 nonwords. All words were unambiguous; their pronunciation was unequivocal, that is, they could be read as a meaningful word in only one way. Nonwords were created by altering one letter of a real word, and could not be read as a meaningful word with any vowel configuration. All stimuli were three to five letters long, and contained two syllables with four to six phonemes. The average number of letters and phonemes was similar for the three types of stimuli.

The words were presented to the subjects for lexical decision or naming. Forty-eight different subjects were tested in each task with the same stimuli. The stimuli were presented in four lag conditions 0, 100, 200, and 300 ms. Each lag was defined by the SOA (stimulus onset asynchrony) between the presentation of the consonants and the vowel marks. Thus, at lag 0 the consonants and the vowel marks were presented simultaneously, at lag 100 the consonants were presented first and the vowel marks were superimposed 100 ms later, etc. Four lists of words were formed: Each list contained 160 stimuli that were composed of 10 high-frequency words, 10 low-frequency words, and 20 nonwords in each of the four lag conditions. The stimuli were rotated across lists by a Latin Square design so that words that appeared in one lag in one list, appeared in another lag in another list, etc. The purpose of this rotation was to test each subject in all lagging conditions while avoiding repetitions within a list. The subjects were randomly assigned to each list and to each experimental condition (lexical decision or naming).

Procedure and apparatus. The procedure and apparatus were identical to those used in the previous experiments. The only difference was that the letters and the vowel marks appeared with the different SOAs. The clock for measuring response times was initiated on each trial with the presentation of the letters, regardless of vowel marks. The subjects were informed of the SOA manipulation but were requested to communicate their lexical decisions or vocal responses as soon as possible, that is, without necessarily waiting for the vowel marks to appear. Each session started with 16 practice trials. The 160 test trials were presented in one block.

Results

Mean RTs for high-frequency words, low-frequency words and nonwords in the different lag conditions for both the lexical decision and the naming tasks are presented in Table 3. Although the response pattern at each of the different lags is important, for the purpose of simplicity the "lag effect" specified in Table 3 reflects the difference between the simultaneous presentation (lag 0) and the longest SOA (lag 300).

The lagging of the vowel information had relatively little effect in the lexical decision task. The presentation of vowels marks 300 ms after the letters delayed subjects' responses by only 25 ms for high-frequency words, and 37 ms for low-frequency words. The lagging of vowels had a somewhat greater effect on the nonwords. Across all lags, the frequency effect was stable, about 85 ms on the average. In contrast to lexical decisions, the effect of lagging the vowel information had a much greater influence on naming. However, the frequency effect was very small at lag 0 and 100.

The statistical significance of these differences was assessed in a three-way ANOVA with the factors of task (lexical decision, naming), stimulus type (high-, low- frequency, nonwords), and lag (0, 100, 200, 300), across subjects and across stimuli. The main effect of task was significant ($F1(1,94) = 4.2$, $MS_e = 70{,}811$, $p<0.04$; $F2(1,157) = 35$, $MS_e = 8320$, $p<0.001$), as was the main effect of stimulus type ($F1(2,188) = 252$, $MSe = 4914$, $p<0.001$; $F2(2,157) = 106$, $MS_e = 13{,}354$, $p<0.001$), and the main effect of lag ($F1(3,282)=98$, $MS_e = 2495$, $p<0.001$; $F2(3,471) = 94$, $MS_e = 2587$, $p<0.001$). Task interacted with stimulus type ($F1(2,188) = 23.1$, $MS_e = 4914$, $p<0.001$; $F2(2,157) = 13$, $MS_e = 8320$, $p<0.001$) and with lag ($F1(3,282) = 13.4$, $MS_e = 2495$, $p<0.001$, $F2(3,471) = 22$, $MS_e = 1387$, $p<0.001$). Lag interacted with stimulus type ($F1(6,564) = 9.3$, $MS_e = 2076$, $p<0.001$; $F2(6,471) = 9$, $MS_e = 2587$, $p<0.001$). The three-way interaction did not reach significance in the subject analysis($F1=1.3$) but was marginally significant in the stimulus analysis ($F2(6,471) = 2.1$, $MS_e = 1387$; $p<0.05$). A Tukey-A post-hoc test revealed that the frequency effect in naming was significant only at the longer lags (200, 300 ms SOA), and not in the shorter lags (0, 100 ms SOA), whereas in lexical decisions the frequency effect was significant at all lags ($p<0.05$).

Discussion

The results of Experiment 3 suggest that the effective cost of delaying vowel marks in lexical decisions in Hebrew is relatively low. A lag of 300 ms in vowel marks presentation resulted in 25 ms difference in response time for high-frequency words and 37 ms for low-frequency words. This outcome suggests that subjects were not inclined to wait for the vowel marks to formulate their lexical decisions; rather, their responses were based on the recognition of the letter cluster. This interpretation converges with previous studies that showed that lexical decisions in Hebrew are not based on a detailed phonological analysis of the printed word, but rely on a fast judgment of the printed word's orthographic familiarity based on visual access to the lexicon (Bentin & Frost, 1987; Frost & Bentin, 1992b). The slightly higher cost of lagging the vowel marks with low-frequency words and the even higher one for nonwords proposes that stimulus familiarity played a role in the decision strategy. This suggests that for words that were less familiar, or for nonwords, subjects were more conservative in their decisions and were inclined to wait longer for the vowel marks, for the possible purpose of obtaining a prelexical phonological code as well. In addition to the lag effect, a strong frequency effect was obtained at all lags. This provides further confirmation of lexical involvement in the task, regardless of the vowel mark presentation.

A very different pattern of results emerged in the naming task. The effective cost of delaying the vowel marks in naming was much higher. The lag effect obtained in naming words was about 70 ms on the average, twice as large as the effect found for lexical decisions. Thus, although the phonologic structure of the unambiguous words could be unequivocally retrieved from the lexicon following visual access (addressed phonology), subjects were more inclined to wait for the vowels to appear in the naming task than in the lexical decision task, presumably in order to generate a prelexical phonologic code. The lag effect for words suggests that subjects combined both prelexical and lexical routines to name words, but the relative use of the prelexical and lexical routines varied as a function of the delay in vowel mark presentation. This lag effect should be first compared to the effect obtained for naming nonwords. Because they cannot be named correctly without the vowel marks, the lag effect for nonwords reflects the overall cost in response time due to a 300 ms delay of vowel mark presentation. This cost is smaller than the lag itself because the articulation program can be initiated as the letters appear, previous to the vowel mark presentation. The results suggest that the cost of lagging the vowel marks by 300 ms is about 150 ms in response time.

Table 3. *Lexical decision and naming RTs and percent errors for high-frequency words (HF), low-frequency words (LF), and nonwords (NW) when vowel marks appear at different lags after letter presentation. Words are unambiguous.*

	Lexical Decision					Naming				
LAG	0	100	200	300	Lag Effect	0	100	200	300	Lag Effect
HF	546 (6%)	551 (6%)	560 (7%)	571 (6%)	25	587 (4%)	608 (3%)	610 (4%)	648 (3%)	62
LF	626 (12%)	642 (10%)	644 (11%)	663 (10%)	37	598 (5%)	624 (3%)	648 (4%)	678 (5%)	80
Frequency effect	80	91	84	92		11	16	38	30	
NW	641 (9%)	664 (8%)	684 (8%)	710 (9%)	69	655 (5%)	700 (7%)	736 (5%)	799 (7%)	144

Thus, the adoption of a pure prelexical strategy of naming words should have resulted in a similar lag effect. The results suggest that this was not the strategy employed in naming words: The lag effect for words was only half the effect obtained for nonwords. However, the lag effect for naming words should be also compared to the lag effect obtained in the lexical decision task. Note that, as in the lexical decision task, subjects did not need the vowel marks (and a prelexical phonologic code) to name the words correctly. Nevertheless, in contrast to lexical decision, they preferred to wait longer for the vowel information. It could be argued that the introduction of nonwords introduced the prelexical strategy in naming. However, since the lag effect for words was much smaller than the lag effect for nonwords, it suggests that a combined use of the assembled and addressed routine was used in pronunciation.

Another possible interpretation could be suggested, however, to account for the lag effect in naming. Because in the naming task subjects are required to produce the correct pronunciation of the printed word, they might have adopted a strategy of waiting for the vowels in order to verify the phonologic structure of the printed word which they generated lexically. By this interpretation, phonology was lexically addressed but verified at a second stage against the delayed vowel information to confirm the lexically retrieved pronunciation. The verification interpretation would regard the difference in naming words and nonwords in this paradigm as the difference between having words verified by the subsequently presented vowels, which is short, and having to construct a pronunciation from the vowel information, which is longer.

Although this interpretation is plausible, it is not entirely supported by the effect of word frequency on response latencies. The suggestion that the combined use of prelexical and lexical phonology varied as a function of the cost of generating a prelexical phonological code is reinforced by examining the frequency effect in naming. In contrast to lexical decision, the frequency effect in naming was overall much smaller, especially at the shorter SOAs. At the 0 ms SOA the frequency effect decreased from 80 ms in lexical decision to 11 in naming. This supports the conclusion that naming with the simultaneous presentation of vowels was mainly prelexical and scarcely involved the mental lexicon. The longer the SOA, the larger the cost of generating a prelexical phonologic code and, consequently, the greater the use of addressed lexical phonology. This is reflected in the increase in the frequency effect at the longer SOAs (38 and 30 ms for 200 and 300 ms SOA, respectively). This pattern stands in sharp contrast to the lexical decision task, which yielded very similar frequency effects at all lags.[2]

EXPERIMENT 4

The aim of Experiment 4 was to examine the effect of lagging the vowel marks when phonologically ambiguous words (heterophonic homographs) are presented for lexical decision or naming. In contrast to unambiguous words, the correct phonological structure of Hebrew heterophonic homographs can be determined unequivocally only by referring to the vowel marks, which specify the correct phonological alternative and consequently the correct meaning. Thus, if the analysis provided in the previous

experiment concerning the cost of lagging the vowel marks is correct, the effect of presenting ambiguous words on naming should be similar to the effect of presenting nonwords. In both cases, subjects would have to rely on the vowel marks for generating a phonological representation necessary for pronunciation. In contrast, lagging the vowel marks of ambiguous words should affect lexical decisions to a much lesser extent. This is because lexical decisions for ambiguous words have been previously shown to be based on the abstract orthographic structure, and occur prior to the process of phonological disambiguation (Bentin & Frost, 1987; Frost & Bentin, 1992a).

In Experiment 4 subjects were presented with letter strings that could be read as two meaningful words, depending on the vowel configuration assigned to the letters. Nonwords were presented as well. The vowel marks were superimposed on the letters at different lags, the same as those employed in Experiment 3. The relative use of orthographic and phonologic coding in the two tasks was again assessed by measuring the effect of lagging the vowel information on lexical decision and naming.

Method

Subjects. One hundred and sixty undergraduate students from the Hebrew University, all native speakers of Hebrew, participated in the experiment for course credit or for payment. Eighty participated in the lexical decision task and 80 in the naming task. None of the subjects had participated in the previous experiment.

Stimuli and design. The words were 40 ambiguous consonant strings each of which represented both a high-frequency and a low-frequency word. The two phonological alternatives were mostly nouns or adjectives that were not semantically or morphologically related. The procedure for assessing the subjective frequencies of the words was similar to the one employed in the previous experiments: Fifty undergraduate students were presented with lists that contained the pointed disambiguated words related to the 40 ambiguous letter strings, and rated the frequency of each word on a 7-point scale. The rated frequencies were averaged across all 50 judges. Each of the 40 homographs that were selected for this study represented two words that differed in their rated frequency by at least 1 point on that scale. As in the previous experiments, the nonwords were constructed by altering one or two letters of meaningful words. The design was similar to that of Experiment 3. However, in order to avoid repetition of the same letter string with different vowel marks, eight lists of words were presented to the subjects instead of four (hence the larger number of subjects). Each list contained only one form (the dominant or the subordinate) of each homograph, at one of the possible four lags, so that each subject was presented with 40 words and 40 nonwords in a list. The stimuli were rotated across lists by a Latin Square design, and consequently each letter string was presented with both vowel mark configurations at all possible lags.

Procedure and apparatus. The procedure and apparatus were identical to those of Experiment 3. Each session started with 16 practice trials, followed by the 80 test trials, which were presented in one block.

Results

Mean RTs for the dominant alternatives, the subordinate alternatives and the nonwords in the different lag conditions for both the lexical decision and the naming tasks are presented in Table 4. As in Experiment 3, the effect of lagging the vowel information had little influence on lexical decision latencies. In fact, the use of ambiguous words reduced the difference between the simultaneous presentation of vowels and their presentation 300 ms after the letters to 21 ms for words on the average, and to only 6 ms for nonwords. A different pattern emerged in the naming task. The effects of lagging the vowel marks on RTs were twice as large as the effects found for unambiguous words in Experiment 3, where there was only one meaningful pronunciation. In fact, the lag effect on ambiguous words was virtually identical with the lag effect on nonwords.

The statistical significance of these differences was assessed in a three-way ANOVA with the factors of task (lexical decision, naming), stimulus type (high-, low- frequency, nonwords), and lag (0, 100, 200, 300), across subjects and across stimuli. The main effect of task was significant ($F1(1,158) = 39$, $MS_e = 114{,}2811$, $p<0.001$; $F2(1,117) = 314$, $MS_e = 7379$, $p<0.001$), as was the main effect of stimulus type ($F1(2,316) = 42$, $MS_e = 7541$, $p<0.001$; $F2(2,117) = 10.5$, $MS_e = 11{,}737$, $p<0.001$), and the main effect of lag ($F1(3,474)=69$, $MS_e = 8459$, $p<0.001$; $F2(3,351) = 113$, $MS_e = 2528$, $p<0.001$). Task interacted with stimulus type ($F1(2,316) = 6.6$, $MS_e = 7541$, $p<0.001$; $F2(2,117) = 5.8$, $MS_e = 7{,}379$, $p<0.001$) and with lag ($F1(3,474) = 46$, $MS_e = 8459$, $p<0.001$, $F2(3,351)=76$, $MS_e = 2616$, $p<0.001$). Lag did not interact with stimulus type ($F1, F2 < 1.0$). The three-way interaction did not reach significance ($F1 = 1.3$, $F2 = 1.4$).

Table 4. *Lexical decision and naming RTs and percent errors for high-frequency words (HF), low-frequency words (LF), and nonwords (NW) when vowel marks appear at different lags after letter presentation. Words are ambiguous.*

	Lexical Decision					Naming				
LAG	0	100	200	300	Lag Effect	0	100	200	300	Lag Effect
Dominant	585 (4%)	608 (5%)	608 (3%)	611 (3%)	26	619 (3%)	650 (2%)	692 (4%)	763 (3%)	142
Subordinate	611 (7%)	620 (5%)	626 (4%)	627 (2%)	16	641 (5%)	699 (4%)	739 (6%)	790 (4%)	150
NW	624 9%)	629 (9%)	635 (10%)	630 (10%)	6	677 (5%)	713 (4%)	755 (8%)	828 (7%)	151

Discussion

The results of Experiment 4 suggest that the cost of delaying the vowel marks for phonologically ambiguous letter strings in the lexical decision task was very low. This outcome supports the conclusions put forward by Frost and Bentin (1992a), suggesting that lexical decisions in Hebrew are based on the recognition of the abstract root or orthographic cluster and do not involve access to a specific word in the phonologic lexicon.

The longest delays of response times due to the lagging of vowel information occurred in the naming task. This was indeed expected. Because the correct pronunciation of phonologically ambiguous words was unequivocally determined only after the presentation of the vowel marks, subjects had to wait for the vowels to appear in order to name those words correctly. Thus, these stimuli provide a baseline for assessing the effect of lagging the vowel marks on naming latencies. This baseline confirms the previous assessment, which was based on the responses to nonwords in Experiment 3, and suggests that, overall, 300 ms in delaying the vowel marks cost 150 ms in response time for articulation if the vowels are necessary for correct pronunciation. Note that, in contrast to the unambiguous words of Experiment 3, the difference in RTs between dominant and subordinate alternatives of ambiguous words cannot reveal the extent of lexical involvement. First, this difference cannot be accurately labeled as a frequency effect. The two phonological alternatives of each homograph differed only in their dominance, and thus could have been both frequent or both nonfrequent. Moreover, there is a qualitative difference between frequency effects that arises from lexical search and the dominance effect that results from the conflict between two phonological alternatives of heterophonic homographs. Thus, the slower RTs of subordinate alternatives, at least in the naming task, were probably due not to a longer lexical search for low-frequency words, but to a change in the articulation program. If subjects first considered pronouncing the dominant alternative after the ambiguous consonants were presented, the later appearance of the vowel marks that specified the subordinate alternative would have caused a change in their articulation plan, resulting in slower RTs.

General Discussion

The present study investigated the relative use of assembled and addressed phonology in naming unpointed and pointed printed Hebrew words. Experiment 1 demonstrated that the lexical status of the stimulus had greater effect in unpointed than in pointed print. Experiment 2 confirmed that semantic priming facilitated naming to a lesser extent in the shallow pointed orthography than in the deeper unpointed orthography, even though nonwords were not included in the stimulus list. Experiment 3 and 4 examined the effect of delaying the vowel mark presentation on lexical decision and naming, in order to assess their contribution and importance in the two tasks. Because the vowel marks allow fast conversion of the graphemic structure into a phonologic representation using prelexical conversion rules, their delay constitutes an experimental manipulation that reflects the cost of assembling phonology from print. The two experiments showed that, although both naming and lexical decision could be

performed without considering the vowel marks, subjects were inclined to use them to a greater extent in the naming task when the cost in response delay was low.

These results provide strong support for the weaker version of the ODH. The disagreement between the alternative view and the ODH revolves around the extent of using assembled phonology in shallow orthographies. It is more or less unanimously accepted that in deep orthographies readers prefer to use the addressed routine in naming. This is because the opacity of the relationship between orthographic structure and phonologic form, which is characteristic of deep orthographies, prevents readers from assembling a prelexical phonological representation through the use of simple GPC rules. Thus, the major debate concerning the validity of the ODH often takes place on the territory of shallow orthographies, in order to show their extensive use of addressed phonology in naming. What lies behind the alternative view, therefore, is the assumption (even axiom) that the use of the addressed routine for naming constitutes the least cognitive effort for *any* reader in *any* alphabetic orthography.

The advantage of examining the validity of the ODH by investigating reading in pointed and unpointed Hebrew is therefore multiple. First, it is well established that Hebrew readers are used to accessing the lexicon by recognizing the word's orthographic structure. The phonologic information needed for pronunciation is then addressed from the lexicon (Frost & Bentin, 1992b). The question of interest, then, is: what reading strategies are adopted by Hebrew readers when they are exposed to a shallower orthography? Do they adopt a prelexical strategy even though it is not the natural strategy for processing most Hebrew reading material? A demonstration of the extensive use of the assembled routine in pointed Hebrew therefore provides strong support for the ODH. If readers of Hebrew prefer the use of the assembled routine, surely habitual readers of shallower orthographies would have a similar preference.

The second advantage in using the two orthographies of Hebrew is that it allows the manipulation of a within-language between-orthography design. This design has a methodological advantage over studies that compare different languages. The interpretation of differences in reading performance between two languages as reflecting subjects' use of pre- vs. post-lexical phonology can be criticized on methodological grounds. The correspondence between orthography and phonology is only one dimension on which two languages differ. English and Serbo-Croatian, for example, differ in grammatical structure and in the size and organization of the lexicon. These confounding factors, it can be argued, may also affect subjects' performance. The comparison of pointed and unpointed orthography in one language, Hebrew, allows these pitfalls to be circumvented.

Taken together, the four experiments suggest that the presentation of vowel marks in Hebrew encourages the reader to generate a prelexical phonologic representation for naming. The use of the assembled routine was detected by examining both frequency and semantic priming effects in naming. Experiment 3 provides an important insight concerning the preference for prelexical phonology. When the vowels appeared simultaneously with the consonants, the frequency effect in naming was small and nonsignificant, again suggesting minimal lexical involvement. Thus the results of this condition replicate the findings of Experiment 1.

Two methodological factors should be considered in interpreting the results of Experiments 3 and 4. The first relates to the experimental demand characteristics of the lagged pointed presentation. It could be argued that this unnatural presentation of printed information encouraged subjects to wait for the vowels to appear and consequently to use them. In fact, why not adopt a strategy of waiting for all possible information to be provided? Although this could be a reasonable solution for efficient performance in these experiments, the results of the lexical decision task make it very unlikely that this simple strategy was adopted by our subjects. In this task subjects were not inclined to wait for the vowel marks. This result was most conspicuous in Experiment 4, in which the lag effect was minimal. This outcome suggests that the lagged presentation did not induce a uniform strategy of vowel mark processing, but that subjects adopted a flexible strategy characterized by a gradual pattern of relying more and more on the vowel marks (prelexical phonology) as a function of the task and of the ambiguity of the stimuli.

The differential effect of lag on ambiguous and unambiguous words in the two tasks suggests that both the assembled and the addressed routines were used for generating phonology from print. On one hand subjects used explicit vowel information employing prelexical transformation rules. This is reflected in the greater effect of lag on naming relative to lexical decision latencies, and in the smaller frequency effect of experiment 3 at the

shorter SOAs. On the other hand, it is clear that the phonologic structure of unambiguous words was generated using the addressed routine as well. This is reflected in the differential effect of lagging the vowel information on naming ambiguous and unambiguous words: There were smaller effects of lag on naming unambiguous than ambiguous words. This confirms that subjects did not wait for the vowel marks in order to pronounce the unambiguous words, as long as they waited in order to pronounce the ambiguous words. The flexibility in using the two routines for naming unambiguous words was affected, however, by the "cost" of the vowel information. When no cost was involved in obtaining the vowel information (simultaneous presentation), the prelexical routine was preferred, as reflected by the small frequency effect. In contrast, when the use of vowel information involved a higher cost, given the delayed presentation of the vowel marks, a gradually greater reliance on the addressed routine was observed. This is well reflected by the larger frequency effect at the longer lags. This pattern stands in sharp contrast to the lexical decision task, in which the frequency effect was stable and very similar across lags.

The second methodological issue to be considered in interpreting the results of Experiments 3 and 4 is the presence of nonwords in the experiments. One factor that has been recently proposed to account for the conflicting results concerning the effect of orthographic depth is the inclusion of nonwords in the stimulus list (Baluch & Besner, 1991). In their study, Baluch and Besner presented native speakers of Persian with opaque and transparent Persian words and showed that differences in semantic facilitation in the naming task appeared only when nonwords were included in the stimulus list. When the nonwords were omitted, no differences in semantic facilitation were found. Baluch and Besner concluded that the inclusion of nonwords in the list encourages subject to adopt a prelexical strategy of generating phonology from print, because the phonologic structure of nonwords cannot be lexically addressed but only prelexically assembled.

The results of Experiment 3 and 4 do not support the view that the nonwords induced a pure prelexical strategy. If this were so, the effects of lag would have been similar for words and for nonwords. Subjects would have waited for the vowels of every stimulus to appear and the lag effect for unambiguous words, ambiguous words, and nonwords would have been similar, about 150 ms. This clearly was not the outcome we obtained. Subjects waited the full 150 ms to pronounce the nonwords and the ambiguous words but waited only half as much time to pronounce the unambiguous words. This confirms a mixed strategy in reading.

However, the argument proposed by Baluch and Besner deserves serious consideration. Not only is its logic compelling, but also the effect of nonwords on reading strategies is well documented in several studies. The crux of the debate is, however, whether this argument can account for the observed effects of orthographic depth found in the various studies described above (e.g., Frost et al., 1987; Katz & Feldman, 1983). The evidence for that claim seems to be much less compelling. There is no argument that nonwords induce a prelexical strategy of reading in all orthographies (e.g., Hawkins, Reicher, & Rogers, 1976; and see McCusker, Hillinger, & Bias, 1981, for a review). However, note that the weak ODH proposes that whatever the effect of nonword inclusion is, it would be different in deep and in shallow orthographies. Thus, evidence for or against the weak ODH could only be supported by studies directly examining the *differential* effect of nonwords in various orthographies. Such a design was indeed employed by Baluch and Besner (1991), but their conclusions hinge on the nonrejection of the null hypothesis. In contrast to their results, several studies have provided clear evidence supporting the weak ODH. For example, Frost et al. (1987) showed that the ratio of nonwords had a dramatically different effect on naming in Hebrew, in English, and in Serbo-Croatian. Different effects of nonwords inclusion on semantic facilitation in naming in the deep English and the shallow Italian, were also reported by Tabossi and Laghi (1992).

The results of Experiment 2 also stand in sharp contrast to Baluch and Besner's conclusions. In Experiment 2 different semantic priming effects were found between pointed and unpointed Hebrew, even though nonwords were not included in the stimulus list. How can this difference be accounted for? A possible explanation for this discrepancy could be related to the strength of the experimental manipulation employed in the two studies. Baluch and Besner used opaque and transparent words within the unpointed-deep Persian writing system. The phonological transparency of their words was due to the inclusions of letters that convey vowels. However, because these letters can also convey consonants in a different context, they may have introduced some ambiguity in the print that is characteristic of

deep orthographies, hereby encouraging the use of the addressed routine. In contrast, the experimental manipulation of using pointed and unpointed Hebrew print is much stronger. Hence, differential effects of semantic facilitation in naming emerged in Hebrew, providing strong support for the weak ODH.

The debate concerning the ODH, however, cannot simply revolve around the interpretation of various findings without adopting a theoretical framework for reading in general. Interestingly, while contrasting the ODH with the alternative view, one might find very similar definitions that capture these conflicting approaches. Proponents of the alternative view would argue that, regardless of the characteristics of their orthography readers in different languages seem to adopt remarkably similar mechanisms in reading (e.g., Besner & Smith, 1992; Seidenberg, 1992). Thus, the alternative view attempts to offer a universal mechanism that portrays a vast communality in the reading process in different languages. Surprisingly, proponents of the ODH take a very similar approach. They suggest a basic mechanism for processing print in all languages, which is finely tuned to the particular structure of every language (Carello et al., 1992). The major discussion is, therefore, what exactly this mechanism is.

The basic assumption of the ODH concerns the role of phonology in reading. It postulates as a basic tenet that all writing systems are phonological in nature and their primary aim is to convey phonologic structures, i.e. words, regardless of the graphemic structure adopted by each system (see De Francis, 1989; Mattingly, 1992, for a discussion). Thus, the extraction of phonologic information from print is the primary goal of the reader, whether skilled or beginner. It is the emphasis upon the role of phonology in the reading process that bears upon the importance of prelexical phonology in print processing. The results of the present study furnish additional support for an increasingly wide corpus of research that provides evidence confirming the role of prelexical phonology in reading. This evidence comes not only from shallow orthographies like Serbo-Croatian (e.g., Feldman & Turvey, 1983; Lukatela & Turvey, 1990), but also from deeper orthographies like English, using a backward masking paradigm (e.g., Perfetti et al., 1988) or a semantic categorization task with pseudohomophonic foils (Van Orden, 1987; Van Orden, Johnston, & Halle, 1988). Recently, Frost and Bentin (1992a) showed that phonologic analysis of printed heterophonic homographs in the even deeper unpointed Hebrew orthography precedes semantic disambiguation. In another study, Frost and Kampf (1993) showed that the two phonologic alternatives of Hebrew heterophonic homographs are automatically activated following the presentation of the ambiguous letter string.

The results of the present study converge with these findings. They provide an opportunity to examine the ODH and the alternative view not in the linguistic environment of shallow orthographies, but of deep orthographies. If prelexical phonology plays a significant role in the reading of pointed Hebrew by readers who are trained to use mainly the addressed routine for phonological analysis, then the plausible conclusion is that, in any orthography, assembled phonology plays a much greater role in reading than the alternative view would assume.

REFERENCES

Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance, 10,* 340-357.

Baluch, B., & Besner, D. (1991). Strategic use of lexical and nonlexical routines in visual word recognition: Evidence from oral reading in Persian. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 644-652.

Bentin, S., Bargai, N., & Katz, L. (1984). Orthographic and phonemic coding for lexical access: Evidence from Hebrew. *Journal of Experimental Psychology: Learning Memory & Cognition, 10,* 353-368.

Bentin, S., & Frost, R. (1987). Processing lexical ambiguity and visual word recognition in a deep orthography. *Memory & Cognition, 15,* 13-23.

Besner, D., & Hildebrant (1987). Orthographic and phonological codes in the oral reading of Japanese Kana. *Journal of Experimental Psychology: Learning Memory & Cognition, 13,* 335-343.

Besner, D., & Smith, M. C. (1992). Basic processes in reading: Is the orthographic depth hypothesis sinking? In R. Frost & L. Katz (Eds.), *Orthography, phonology, morphology, and meaning* (pp. 45-66). Amsterdam: Advances in Psychology, Elsevier Science Publishers.

Besner, D., Patterson, K., Lee, L., & Hildebrant N. (1992). Two forms of Japanese Kana: Phonologically but NOT orthographically interchangeable. *Journal of Experimental Psychology: Learning Memory & Cognition.*

Carello, C., Lukatela, G., & Turvey, M. T. (1988). Rapid naming is affected by association but not syntax. *Memory & Cognition, 16,* 187-195.

Carello, C., Turvey, M. T., & Lukatela, G. (1992). Can theories of word recognition remain stubbornly nonphonological? In R. Frost & L. Katz (Eds.) *Orthography, phonology, morphology, and meaning* (pp. 211-226). Amsterdam: Advances in Psychology, Elsevier Science Publishers.

Colombo, L., & Tabossi, P. (1992). Strategies and stress assignment: Evidence from a shallow orthography. In R. Frost & L. Katz (Eds.), *Orthography, phonology, morphology, and meaning* (pp. 319-340). Amsterdam: Advances in Psychology, Elsevier Science Publishers.

DeFrancis, J. (1989). *Visible speech: The diverse oneness of writing systems*. Honolulu: University of Hawaii Press.

Feldman, L. B., & Turvey, M. T. (1983). Word recognition in Serbo-Croatian is phonologically analytic. *Journal of Experimental Psychology: Human Perception and Performance, 9*, 228-298.

Frost, R. (1992). Orthography and phonology: The psychological reality of orthographic depth. In M. Noonan, P. Downing, & S. Lima (Eds.), *The linguistics of literacy* (pp. 255-274). Amsterdam/Philadelphia: John Benjamins Publishing CO.

Frost, R., Katz, L., & Bentin, S. (1987). Strategies for visual word recognition and orthographical depth: A multilingual comparison. *Journal of Experimental Psychology: Human Perception and Performance, 13*, 104-115.

Frost, R., & Katz, L. (1989). Orthographic depth and the interaction of visual and auditory processing in word recognition. *Memory & Cognition, 17*, 302-311.

Frost, R., & Bentin, S. (1992a). Processing phonological and semantic ambiguity: Evidence from semantic priming at different SOAs. *Journal of Experimental Psychology: Learning, Memory, and Cognition. 18*, 58-68.

Frost, R., & Bentin, S. (1992b). Reading consonants and guessing vowels: Visual word recognition in Hebrew orthography. In R. Frost & L. Katz (Eds.), *Orthography, phonology, morphology, and meaning* (pp. 27-44). Amsterdam: Advances in Psychology, Elsevier Science Publishers.

Frost, R., & Kampf, M. (1993). Phonetic recoding of phonologically ambiguous printed words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 1-11.

Hawkins, H. L., Reicher, G. M., & Rogers, M. (1976). Flexible coding in word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 2*, 235-242.

Katz, L. & Feldman L. B. (1981). Linguistic coding in word recognition. In: A.M. Lesgold & C.A. Perfetti (Eds.), *Interactive processes in reading*. Hillsdale, NJ: Erlbaum.

Katz, L., & Feldman L. B. (1983). Relation between pronunciation and recognition of printed words in deep and shallow orthographies. *Journal of Experimental Psychology: Learning, Memory, and Cognition., 9*, 157-166.

Katz, L., & Frost, R. (1992). Reading in different orthographies: the orthographic depth hypothesis. In: R. Frost & L. Katz (Eds.), *Orthography, phonology, morphology, and meaning* (pp. 67-84). Advances in Psychology, Elsevier, North-Holland.

Klima, E. S. (1972). How alphabets might reflect language. In F. Kavanagh & I. G. Mattingly (Eds.), *Language by ear and by eye*. The MIT Press, Cambridge, Massachusetts, and London, England.

Laudanna, A., & Caramazza, A. (1992). Morpho-lexical representations and reading. Paper presented at the fifth conference of the European Society for Cognitive Psychology, Paris, France.

Liberman, I. Y., Liberman, A. M., Mattingly, I. G., & Shankweiler, D. (1980). Orthography and the beginning reader. In J. F. Kavanagh & R. L. Venezky (Eds.), *Orthography, reading, and dyslexia*. Austin, TX: Pro-Ed.

Lukatela, G., Popadić, D., Ognjenović, P., & Turvey, M. T. (1980). Lexical decision in a phonologically shallow orthography. *Memory & Cognition, 8*, 415-423.

Lukatela, G., Feldman, L. B., Turvey, M. T., Carello, C., & Katz, L. (1989). Context effects in bi-alphabetical word perception. *Journal of Memory and Language, 28*, 214-236.

Lukatela, G., & Turvey, M. T. (1990). Automatic and prelexical computation of phonology in visual word identification. *European Journal of Cognitive Psychology, 2*, 325-343.

Lupker, J. S. (1984). Semantic priming without association: A second look. *Journal of Verbal Learning and Verbal Behavior, 23*, 709-733.

McCusker, L. X., Hillinger, M. L., & Bias, R. G. (1981). Phonologic recoding and reading. *Psychological Bulletin, 89*, 217-245.

Mattingly, I. G. (1992). Linguistic awareness and orthographic form. In R. Frost & L. Katz (Eds.), *Orthography, phonology, morphology, and meaning* (pp. 11-26). Amsterdam: Advances in Psychology, Elsevier Science Publishers.

Neely, J. H. (1991). Semantic priming effects in visual word recognition. In D. Besner & G. W. Humphreys (Eds.), *Basic processes in reading: Visual word recognition*. Hillsdale, NJ: Erlbaum.

Perfetti, C. A., Bell, L. C., & Delaney, S. M. (1988). Automatic (prelexical) phonetic activation in silent word reading: Evidence from backward masking. *Journal of Memory and Language, 27*, 59-70.

Perfetti, C.A., Zhang, S., & Berent, I. (1992). Reading in English and Chinese: Evidence for a universal phonological principle. In R. Frost & L. Katz (Eds.), *Orthography, phonology, morphology, and meaning* (pp. 227-248). Amsterdam: Advances in Psychology, Elsevier Science Publishers.

Scheerer, E. (1986). Orthography and lexical access. In G. Augst (Ed.), *New trend in graphemics and orthography*. (pp. 262-286). Berlin: De Gruyter.

Sebastián-Gallés, N. (1991). Reading by analogy in a shallow orthography. *Journal of Experimental Psychology: Human Perception and Performance, 17*, 471-477.

Seidenberg, M. S. (1985). The time course of phonological code activation in two writing systems. *Cognition, 19*, 1-30.

Seidenberg, M. S. (1992). Beyond orthographic depth in reading: Equitable division of labor. In R. Frost & L. Katz (Eds.), *Orthography, phonology, morphology, and meaning* (pp. 85-118). Amsterdam: Advances in Psychology, Elsevier Science Publishers.

Seidenberg, M. S. & Vidanović, S. (1985). Word recognition in Serbo-Croatian and English: Do they differ? Paper presented at the Twenty-fifth Annual Meeting of the Psychonomic Society, Boston.

Tabossi, P., & Laghi, L. (1992). Semantic priming in the pronunciation of words in two writing systems: Italian and English. *Memory & Cognition, 20*, 303-313.

Turvey, M. T., Feldman. L. B., & Lukatela, G. (1984). The Serbo-Croatian orthography constrains the reader to a phonologically analytic strategy. In L. Henderson (Ed.), *Orthographies and reading: Perspectives from cognitive psychology, neuropsychology, and linguistics* (pp. 81-89). Hillsdale, NJ: Erlbaum.

Van Orden, G. C. (1987). A ROWS is a ROSE: Spelling, sound and reading. *Memory & Cognition, 15*, 181-198.

Van Orden, G. C., Johnston, J. C., & Halle, B. L. (1988). Word identification in reading proceeds from spelling to sound to meaning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 371-386.

FOOTNOTES

Journal of Experimental Psychology: Learning, Memory, and Cognition, in press.

[1] The error rates presented in this study should be evaluated with care because the criteria for assessing an error in pointed and unpointed print are different. While in unpointed print any pronunciation consistent with the consonants only is considered correct, in pointed print any pronunciation that is inconsistent with the explicit vowel information is considered an error. This is of special significance while evaluating the nonwords data. In unpointed print subjects have much greater flexibility in naming nonwords than in pointed print, hence the larger error rates in the pointed condition.

[2] But note that the effect of delaying vowels on naming latencies does not linearly decrease with lag, as should be predicted by the increase of frequency effect. In order to obtain a more ordered relationship more lag conditions should have been used.

Haskins Laboratories Status Report on Speech Research
1993, SR-113, 171-196

Relational Invariance of Expressive Microstructure across Global Tempo Changes in Music Performance: An Exploratory Study*

Bruno H. Repp

This study addressed the question of whether the expressive microstructure of a music performance remains relationally invariant across moderate (musically acceptable) changes in tempo. Two pianists played Schumann's "Träumerei" three times at each of three tempi on a digital piano, and the performance data were recorded in MIDI format. In a perceptual test, musically trained listeners attempted to distinguish the original performances from performances that had been artificially speeded up or slowed down to the same overall duration. Accuracy in this task was barely above chance, suggesting that relational invariance was largely preserved. Subsequent analysis of the MIDI data confirmed that each pianist's characteristic timing patterns were highly similar across the three tempi, although there were statistically significant deviations from perfect relational invariance. The timing of (relatively slow) grace notes seemed relationally invariant, but selective examination of other detailed temporal features (chord asynchrony, tone overlap, pedal timing) revealed no systematic scaling with tempo. Finally, although the intensity profile seemed unaffected by tempo, a slight overall increase in intensity with tempo was observed. Effects of musical structure on expressive microstructure were large and pervasive at all levels, as were individual differences between the two pianists. For the specific composition and range of tempi considered here, these results suggest that major (cognitively controlled) temporal and dynamic features of a performance change roughly in proportion with tempo, whereas minor features tend to be governed by tempo-independent motoric constraints.

INTRODUCTION

When different artists perform the same musical composition, they often choose very different tempi. Even though each artist may be convinced that his or her tempo is "right," and even though the composer may have prescribed a specific tempo in the score, there is in fact a range of acceptable tempi for any composition played on conventional instruments. The American composer Ned Rorem has expressed this well:

> Tempos vary with generations like the rapidity of language. Music's velocity has less organic import than its phraseology and rhythmic qualities; what counts in performance is the artistry of phrase and beat within a tempo. ... [The composer's] Tempo indication is not creation, but an afterthought related to performance. Naturally an inherently fast piece must be played fast, a slow one slow—but just to what extent is a decision for players. (Rorem, 1983, p. 326.)

Guttmann (1932) measured the durations of a large number of orchestral performances and found substantial tempo variation across different conductors, for the same composition. In extreme cases, the fastest observed performance was about 30% shorter than the slowest one. In two recent studies, Repp (1990, 1992) compared performances by famous pianists of solo pieces by

This research was made possible by the generosity of Haskins Laboratories (Carol A. Fowler, president). Additional support came from NIH BRSG Grant RR05596 to the Laboratories. I am grateful to LPH for her patient participation in this study, and to Peter Desain and Henkjan Honing for helpful comments on an earlier version of the manuscript.

Beethoven and Schumann and in each case found a considerable range of tempi. The ratio of the extreme tempi was approximately 1:1.6 in each case, corresponding to a 37% difference in performance duration. Tempo differences among repeated performances by the same artist tend to be smaller but may also be considerable (Guttmann, 1932; Repp, 1992), notwithstanding the observation of remarkable constancy for some artists or ensembles (Clynes and Walker, 1986).

Tempo varies not only between but also within performances. Even though no tempo changes may have been prescribed by the composer in the score, a sensitive performer will continuously modulate the timing according to the structural and expressive requirements of the music. This temporal modulation forms part of the "expressive microstructure" of a performance. Naturally, it is more pronounced in slow, "expressive" than in fast, "motoric" pieces. The problem of determining the global or baseline tempo of a highly modulated performance is addressed elsewhere (Repp, in press). In this article the question of interest is whether overall tempo interacts with the pattern of expressive modulations.

The null hypothesis to be tested is that expressive microstructure (which on the piano includes not only successive tone onset timing but also simultaneous tone onset asynchronies, successive tone overlap or separation, pedal timing, and successive as well as simultaneous intensity relationships) is *relationally invariant* across changes in tempo. What this implies is that a change in tempo amounts to multiplying all temporal intervals by a constant, so that all their relationships (ratios) remain intact. Nontemporal properties of performance (i.e., tone intensities) are likewise predicted to remain relationally invariant.

Relational invariance (also called "proportional duration") is a key concept in studies of motor behavior (see Gentner, 1987; Heuer, 1991; Viviani and Laissard, 1991). It has been used as an indicator of the existence of a "generalized motor program" (Schmidt, 1975) having a variable rate parameter. Many activities have been examined from that perspective, and relational invariance has commonly been observed, even though stringent statistical tests may show significant deviations from strict proportionality (see Heuer, 1991). The deviations may result from the combination of more central and more peripheral sources of variability. Heuer (1991) conjectures that relational invariance is most likely to obtain when the motor behavior is "natural" (i.e., conforms to peripheral constraints) and "self-selected" (i.e., not imposed by the experimenter). These conditions certainly apply to artistic music performance, especially when (as in the present study) it is slow and expressive, so that peripheral factors (i.e., technical difficulties) are minimized. Nevertheless, even such a performance has aspects (e.g., the relative asynchrony of chord tones or the relative overlap of *legato* tones on the piano) that are subject to peripheral constraints imposed by fingering, the spatial distance between keys, and the limits of fine motor control. The question of relational invariance thus may be asked at several levels of detail, and the answers may differ accordingly. In sheer complexity and degree of precision, expert musical performance exceeds almost any other task that may be investigated from a motor control viewpoint (see Shaffer, 1980, 1981, 1984).

Few previous studies have looked into the question of relational invariance in music performance. Deviations from relational invariance seem likely when the tempo changes are large. Handel (1986) has pointed out that changes in tempo may cause rhythmic reorganization: A doubling or halving of tempo often results in a change of the level at which the primary beat (the "tactus") is perceived; this changes the rhythmic character of the piece, with likely consequences for performance microstructure. Very fast tempi may lead to technical problems whereas very slow tempi may lead to a loss of coherence. These kinds of issues may be profitably investigated with isolated rhythmic patterns, scales, and the like. When it comes to the artistic interpretation of serious compositions, however, such dramatic differences in tempo are rare. The tempi chosen cannot stray too far from established norms, or else the performance will be perceived as deviant and stylistically inappropriate. In this study, therefore, we will be concerned only with moderate tempo variations that are within the range of aesthetic acceptability for the composition chosen. Still, as indicated above, this range is wide enough to make the question of relational invariance nontrivial.

Since different artists' performances of the same music differ substantially in their expressive microstructure ("interpretation") as well as in overall tempo, it is virtually impossible to determine the relationship between these two variables by comparing recordings of different artists. Although many studies have demonstrated that individual performers are

remarkably consistent in reproducing their expressive timing patterns across repeated performances of the same music, these performances are generally also very similar in tempo (e.g., Seashore, 1938; Shaffer, 1984; Shaffer, Clarke, and Todd, 1985; Repp, 1990, 1992). Where substantial differences in tempo do occur, underlying changes in "interpretation" (i.e., in the cognitive structure underlying the performance) cannot be ruled out (e.g., Shaffer, 1992; Repp, 1992). In other words, the change in tempo may have been caused by a change in interpretation, rather than the other way around. Clearly, artists can change their interpretation at will, and with it the tempo of a performance. The question posed in the present study was whether a change in tempo *necessitates* a change in expressive microstructure. Therefore, an experimental approach was taken in which the same artist was asked to play a piece at different tempi without deliberately changing the interpretation. The resulting performance data were analyzed and compared. This approach was supplemented by a perceptual test in which the tempo of recorded performances was artificially modified and the result was judged by musically trained listeners.

One previous study that directly addressed the hypothesis tested here was conducted by Clarke (1982), following preliminary but inconclusive observations by Michon (1974). Clarke examined selected sections of performances of the highly repetitive piano piece, "Vexations," by Erik Satie, played by two pianists who had been instructed to vary the tempo. The timing patterns (tone inter-onset intervals, or IOIs) of the same music at six different tempi were compared, and a significant interaction with tempo was found. The interaction was in part due to a narrowing of the tempo range in the course of the excerpt, but a significant effect remained after the tempo drift was removed statistically. Subsequent inspection of the data suggested to Clarke that the pianists had changed their interpretation of the grouping structure across tempi. At least one pianist produced more groups at the slower tempo, where group boundaries were identified by a temporary slowing down. Clarke also argued that group boundaries that were maintained tended to be emphasized more at the slower tempi.

While Clarke's interpretation is intuitively plausible, the changes in the timing profiles across tempi were quite small and do not suggest different structural interpretations to this reader. Although there was more temporal modulation at the slower tempi, this is consistent with the hypothesis of relational invariance and may have accounted for the interaction with tempo. Such an interaction might not have appeared in log-transformed data.[1] Finally, even if it were true that the pianists changed their interpretations along with tempo, this may have reflected the extremely boring and repetitive nature of the music which, as the title says, is a deliberate harassment of the performer. Therefore, the generality of Clarke's conclusions is not clear. A very casual report of a follow-up study with a movement from a Clementi Sonatina (Clarke, 1985) does not change the picture significantly.

In a study of piano performance at a simpler level, MacKenzie and Van Eerd (1990) found highly significant changes in tone IOIs as a function of tempo. Their pianists' task was to play scales as evenly as possible, and the range of tempi was very large. The IOI profiles observed were not due to musical structure or expressive intent but to fingering; they became more pronounced as tempo increased. These kinds of motor constraints, though they become important in fast and technically difficult pieces, are likely to play only a minor role in slow, expressive performance, as studied here.

The immediate stimulus for the present study was provided by some informal observations reported by Desain and Honing (1992a). They recorded a pianist playing a tune by Paisiello (the theme of Beethoven's variations on "Nel cor piú non mi sento") at two different tempi, M.M.60 and 90, and observed that the temporal microstructure differed considerably between the two performances. They also used a MIDI sequencer to speed up the slow performance and to slow down the fast performance, and in each case they found that the altered performance sounded unnatural compared to the original performance having the same tempo. On closer examination, the major differences between the two performances seemed to lie in the timing of grace notes, in the articulation of staccato notes, and in the "spread" (onset asynchrony) of chords, though the overall timing profiles also looked quite different.

In a subsequent paper, Desain and Honing (1992b) developed a computer model for implementing changes in expressive timing. They distinguished four types of "musical objects": sequences of tones, chords, *appoggiature*, and *acciaccature*, the last two referring to types of grace notes. Each type of musical object is subject to different temporal transformation rules, which are applied successively within a hierarchical

representation of the musical structure. In the first instance, these transformations represent changes along a continuum of degrees of expressiveness (i.e., deviations from mechanical exactitude) within a fixed temporal frame. However, the temporal changes within a larger unit propagate down to smaller units, effectively altering their tempo. Judging from their Figure 5, Desain and Honing propose that tone onsets within melodic sequences are stretched or shrunk proportionally as tempo changes, whereas the other three types of structures remain temporally constant or get "truncated" (in the case of chord asynchrony). Thus, these authors seem to suggest that relational invariance across changes in tempo does hold in a sequence of melody tones, but not in chords or ornaments. As to articulation (i.e., the overlap or separation of successive tones), they discuss three alternative transformations, without committing themselves. Their model is a system for implementing different types of rules, not a theory of what these rules might be. Nevertheless, their examples reflect some of their earlier informal observations.

The present study investigated not only timing patterns but also tone intensities, another important dimension of expressive performance that has received much less attention in research so far. Todd (1992) has pointed out that an increase of tempo often goes along with an increase in intensity, which leads to the hypothesis that the overall dynamic level of a performance may be affected by a tempo change. MacKenzie and Van Eerd (1990) found such an increase in key press velocity with tempo in their study of pianists playing scales. It is not clear whether any changes in intensity relationships should be expected across tempi, however. MacKenzie and Van Eerd found a subtle interaction, but the differences were not expressively motivated. The hypothesis investigated here was that intensity microstructure would remain relationally invariant across tempo changes, but that pianists might play louder overall at faster tempi. (If this latter effect were absent, the intensity pattern would be absolutely invariant.)

The present study aimed at providing some preliminary data bearing on these issues. The data were limited in so far as they came from a single musical composition played by two pianists, but they included a number of different measurable aspects of performance. The composition, Robert Schumann's "Traumerei," was selected because its expressive timing characteristics and acceptable tempo range were known from Repp's (1992) detailed analysis of 28 expert performances, and also because it is a slow, interpretively demanding, but technically not very challenging piece. It contains no *staccato* notes or ornaments, apart from a few expressive grace notes. Thus the main question was whether the expressive timing of the melody tones would or would not scale proportionally with changes in tempo. However, the tempo scaling of other performance aspects (grace notes, chord asynchrony, tone overlap, pedaling, dynamics) was of nearly equal interest. Based on the limited observations summarized above, it was expected that the overall timing and intensity profiles would exhibit relational invariance across tempo changes, whereas this might not be true for the more detailed temporal features. In the course of the detailed analyses, there were opportunities to make new (and confirm old) observations about the relation of musical structure and performance microstructure, some of which will be discussed briefly in order to characterize the nature of the patterns whose relational (non)invariance was investigated. A detailed investigation of structural factors is beyond the scope of this paper.

In addition to these performance measurements, a perceptual test was carried out along the lines explored by Desain and Honing (1992a), to determine whether artificially speeded-up or slowed-down performances would sound odd compared to original performances having the same overall tempo. The results of that test will be described first, followed by the performance analyses.

Methods

The music

The score of Schumann's "Träumerei" is shown in Figure 1; its layout highlights the hierarchical structure of the piece. The composition comprises three 8-bar sections, each of which contains two 4-bar phrases which in turn are composed of shorter melodic gestures. The first section (bars 1-8) is repeated in performance. Each of the six phrases starts with an upbeat that (except for the one at the very beginning) overlaps with the end of the preceding phrase. Phrases 1 and 5 are identical and similar to phrase 6, whereas phrases 2, 3, and 4 are more similar to each other than to phrases 1, 5, and 6. Thus there are two phrase types here. Vertically, the composition has a four-voice polyphonic structure. For a more detailed analysis, see Repp (1992).

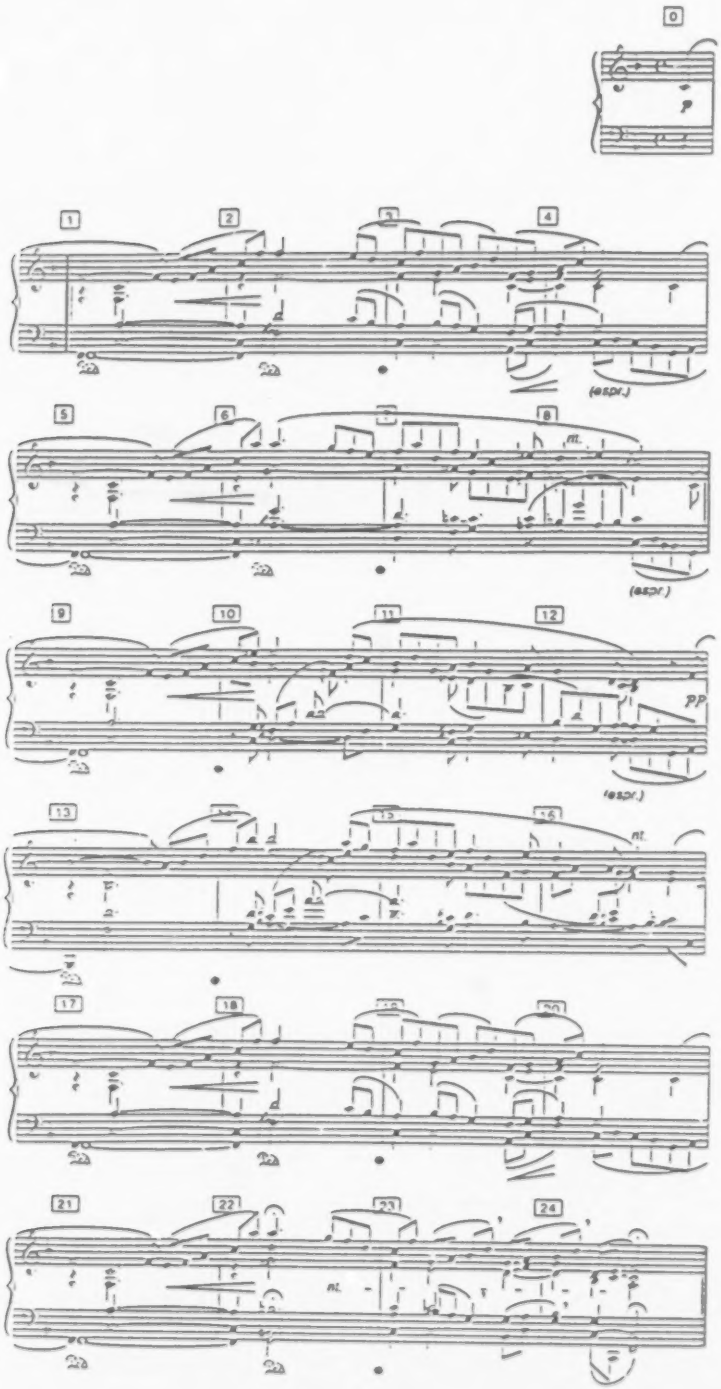


Figure 1. **Score of Schumann's "Träumerei," prepared with MusicProse software following the Clara Schumann edition (Breitkopf & Härtel). Minor discrepancies are due to software limitations. The layout of the computer score highlights parallel musical structures.**

Performers

The two performers were LPH, a professional pianist in her mid-thirties, and BHR (the author), a serious amateur in his late forties. Both were thoroughly familiar with Schumann's "Träumerei" and had played it many times in the past.

Recording procedure

The instrument was a Roland RD250S digital piano with weighted keys and sustain pedal switch, connected to an IBM-compatible microcomputer running the Forte sequencing program. Synthetic "Piano 1" sound was used. The pianists played from the score and monitored the sound over earphones. The computer recorded each performance in MIDI format, including the onset time, offset time, and velocity of each key depression, as well as pedal on and off times. The temporal resolution was 5 ms.

Each pianist was recorded in a separate session. After some practice on the instrument, (s)he played the full piece (including the repeat of the first 8 bars) at her/his preferred tempo. Afterwards, (s)he listened to the beginning of the recorded performance and set a Franz LM-FB-4 metronome to what (s)he believed to be the basic tempo of the performance. The settings chosen by LPH and BHR were M.M.63 and 66, respectively. Subsequently, each pianist performed the piece at a slower and at a faster tempo. These tempi were chosen by the experimenter to be M.M.54 and 72 for LPH and M.M.56 and 76 for BHR. All these tempi were within the range observed in commercially recorded performances by famous pianists (Repp, 1992; in press). The desired tempo was indicated by the metronome before each performance; the metronome was turned off before playing started. The pianists' intention was to play as naturally and expressively as possible at each tempo; no conscious effort was made either to maintain or to change the interpretation across tempi.

Each pianist then repeated the cycle of three performances until three good recordings had been obtained at each tempo.[2] In these repeats, the medium (preferred) tempo was also cued by metronome. LPH's first performance was excluded because it showed signs of her still getting accustomed to the instrument, and an extra medium-tempo performance was added at the end of the session. BHR actually recorded five performances at each tempo, two in one session and three in another. Only the performances from the second session were used. Thus neither pianist's first performance was included in her/his final set of 9 performances.

The MIDI data files were converted into text files and imported into a data analysis and graphics program (DeltaGraph Professional) for further processing. Tone interonset intervals (IOIs) in milliseconds were obtained by computing differences among successive tone onset times. Tone intensities were expressed in terms of raw MIDI velocities ranging from 0 to 127. In the middle range (which accommodated virtually all melody tones), a difference of 4 MIDI velocity units corresponds to a 1 dB difference in peak rms sound level (Repp, 1993).

Perceptual test

Stimuli. The purpose of the perceptual test was to determine whether performances whose tempo has been artificially altered sound noticeably more awkward than unaltered performances. To reduce the length of the test, each performance was truncated after the third beat of bar 8; the cadential chord at that point was extended in duration to convey finality. Moreover, only the first two of each pianist's three performances at each tempo were used. From each of these 12 truncated original performances, two modified performances were generated by artificially speeding up or slowing down its tempo, such that its total duration matched that of an original performance at a different tempo by the same pianist. Thus, for example, LPH's first slow performance was speeded up to match the duration of the first medium performance, and was speeded up further to match the first fast performance. The tempo modification was achieved by changing the "metronome" setting in the Forte sequencing program, relative to the baseline setting of M.M.100 during recording. (This setting was unrelated to the tempo of the recorded performance but determined the temporal resolution.)

Each of the 12 original performances was then paired with each of two duration-matched modified performances, resulting in 24 pairs. The original performance was first in half the pairs and second in the other half, in a counterbalanced fashion. The pairs were arranged in a random order and were recorded electronically onto digital cassette tape. Performances within a pair were separated by about 2 s of silence, whereas about 5 s of silence intervened between pairs. The test lasted about half an hour.

Subjects and procedure. Nine professional-level pianists served as paid subjects. All but one were graduate students at the Yale School of Music, seven of them for a master's degree in piano performance and one for a Ph.D. degree in

composition. They were tested individually in a quiet room. The test tape was played back at a comfortable intensity over Sennheiser HD420SL earphones. The subjects were fully informed about the nature and manipulation of the stimuli, and they were asked to identify the original performance in each pair by writing down "1" or "2," relying basically on which performance sounded "better" to them.

Results and Discussion

Perceptual test

Average performance in the perceptual test was 55.6% correct. This was not significantly above chance across subjects [$F(1,8) = 2.91$, $p < .13$], though it was marginally significant across pairs of items [$F(1,11) = 5.08$, $p < .05$].[3] There was no significant difference between the scores for the two pianists' performances (though BHR's were slightly easier to discriminate) or for different types of comparisons. These results indicate that the tempo transformations did little damage to the expressive quality of the performances, and hence that there were probably no gross deviations from relational invariance, at least in the first 8 bars.

Timing patterns at the preferred tempo

The performance aspect of greatest interest was the pattern of IOIs (the timing pattern or profile). Before examining the effects of changes in tempo on this aspect of temporal microstructure, however, it seemed wise to look at the timing profiles of the performances at the pianists' preferred (medium) tempi, to confirm that they were meaningful, representative, and reliable. This seemed particularly important in view of the fact that an electronic instrument had been used.

Figure 2 plots the onset timing patterns averaged across the three medium-tempo performances of each pianist. The format of this figure is identical to that of Figure 3 in Repp (1992), which shows the Geometric Average timing pattern of 28 different performances by 24 Famous Pianists (GAFP for short). The timing patterns of three related phrases are superimposed in each panel. (The two renditions of bars 1-8 were first averaged.) Interonset intervals longer than a nominal eighth-note are shown as "plateaus" of equal eighth-note intervals.

The two pianists' average medium-tempo performances resembled both each other and the GAFP timing pattern. The correlations of the complete log-transformed IOI profiles (N=254) were 0.85 between LPH and BHR, 0.88 between LPH and GAFP, and 0.94 between BHR and GAFP. Of course, these overall correlations were dominated by the major excursions in the timing profiles. A measure of similarity at a more detailed level was obtained by computing correlations only over the initial 8 bars, which did not contain any extreme *ritardandi*. These correlations (N=130) were 0.64 ($p < .0001$) between LPH and BHR, 0.77 between LPH and GAFP, and 0.83 between BHR and GAFP.[4] Clearly, both pianists' performances were reasonably representative in the sense that they resembled the average expert profile, even though they were only moderately similar to each other at a detailed quantitative level.

Each pianist also showed high individual consistency in terms of timing patterns both within and between the three performances at her/his preferred tempo. Thus the average correlation among the three complete medium-tempo performances was 0.91 for LPH and 0.95 for BHR. For the initial 8 bars the corresponding correlations were 0.90 for both LPH and BHR. Needless to say, the timing patterns of the two renditions of bars 1-8 within each performance were also extremely similar.

Furthermore, as can be seen in Figure 2, both LPH and BHR produced very similar timing profiles for the identical phrases in bars 1-4 and 17-20 (left-hand panels); LPH played the whole phrase slower in bars 17-20 whereas BHR slowed down only in bar 17. Both pianists also showed similar timing patterns for the analogous phrases in bars 9-12 and 13-16 (right-hand panels), which diverged substantially only at the end, due to the more pronounced *ritardando* in bar 16, the end of the middle section. The structurally similar phrase in bars 5-8 (right-hand panels) again showed a similar pattern; apart from slight differences in overall tempo, major deviations from the pattern of the other two phrases occurred only in the second bar, where the music in fact diverges (cf. Figure 1), and also in the fourth bar for BHR, who emphasized the cadence in the soprano voice and treated the phrase-final bass voice as more transitional than did LPH. Finally, bars 21-24 (left-hand panels) retained the qualitative pattern of bars 1-4 but included substantial lengthening due to the *fermata* and the final grand *ritardando*.[5]

Many additional comments could be made about the detailed pattern of temporal variations and their relation to the musical structure, but such a discussion may be found in Repp (1992) and need not be repeated here. Instead, we will focus now on the comparison among the three tempo conditions.

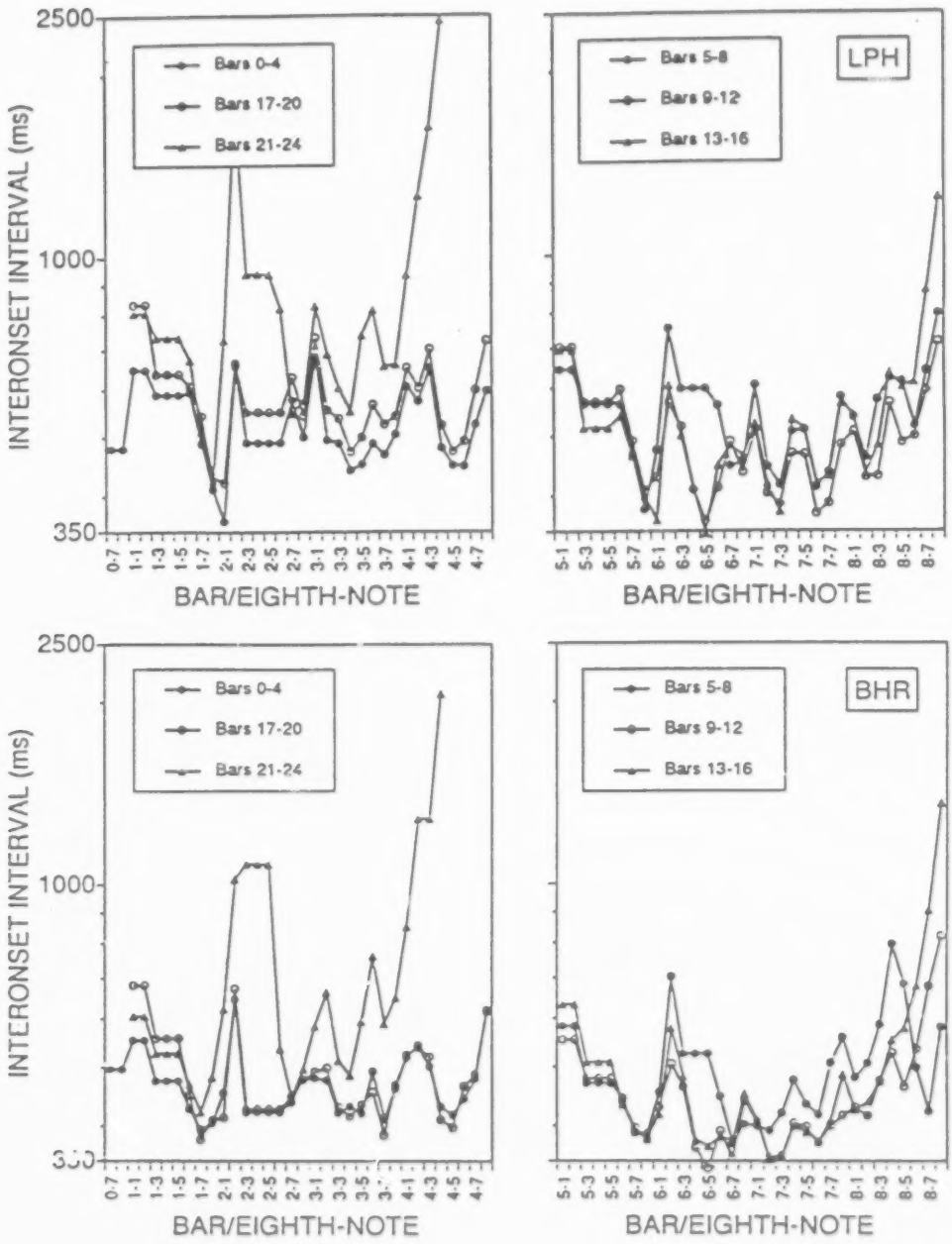


Figure 2. **Eighth-note IOIs as a function of metric distance, averaged across the three medium-tempo performances of each pianist (top: LPH; bottom: BHR). Structurally identical or similar phrases are overlaid in left- and right-hand panels.**

Effects of tempo on global timing patterns

A first question was whether the pianists were as reliable in their timing patterns in the slow and fast conditions as at the medium tempo. After all, they had to play at tempi they would not have chosen themselves. The average overall between-performance correlations within tempo categories for LPH were 0.88 (slow) and 0.91 (fast), as compared to 0.91 (medium); and for BHR they were 0.95 (slow) and 0.94 (fast), as compared to 0.95 (medium). Thus the pianists were just about as reliable in their gross timing patterns at unfamiliar tempi as at the familiar tempo. Computed over bars 1-8 only, the correlations were 0.83 (slow) and 0.86 (fast) versus 0.90 (medium) for LPH, and 0.89 (slow) and 0.86 (fast) versus 0.90 (medium) for BHR. At this more detailed level, then, there is an indication that the pianists were slightly more consistent when they played at their preferred tempo. Clearly, however, they produced highly replicable timing patterns even at novel tempi.

We come now to the crucial question: Were the timing profiles at the slow and fast tempi similar to those at the medium tempo? First, the correlational evidence: The average of the 27 between-tempo correlations (3 x 3 for each of 3 pairs of tempo categories) across entire performances for LPH was 0.90, which was the same as her average within-tempo correlation. For BHR, the analogous correlations were 0.94 and 0.95, respectively. At the more detailed level of bars 1-8, the correlations were both 0.86 for LPH, and 0.87 versus 0.88 for BHR. Thus, between-tempo correlations were essentially as high as within-tempo correlations.

If, as predicted by the relational invariance hypothesis, the intervals at different tempi are proportional, then the average timing profiles should be parallel on a logarithmic scale. This is shown for the first rendition of bars 0-8 in Figure 3. There is indeed a high degree of parallelism between these functions; the few local deviations do not seem to follow an interpretable pattern. These data thus seem consistent with the hypothesis of a multiplicative rate parameter (which becomes additive on a logarithmic scale).

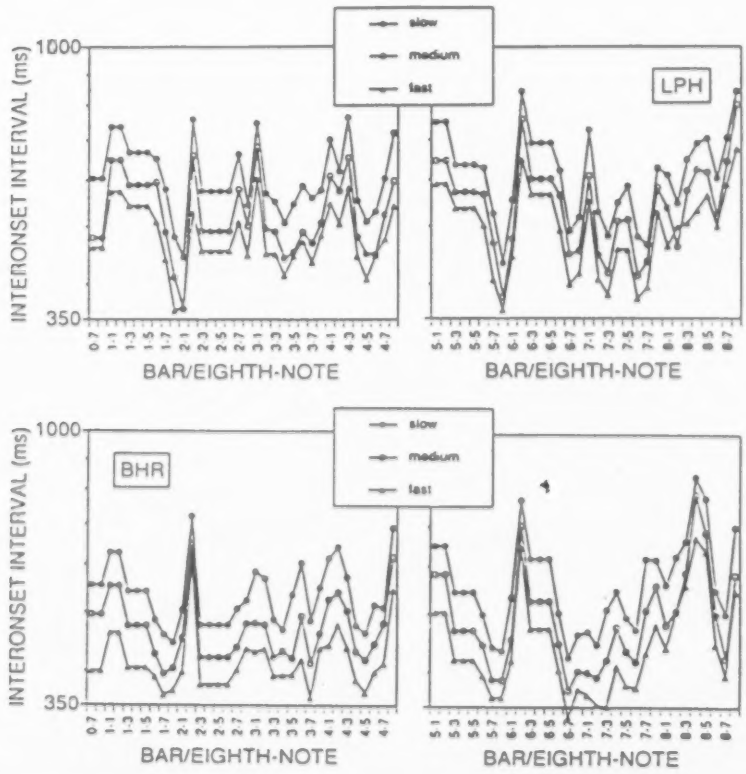


Figure 3. **Eighth-note timing profiles for the initial 8 bars (first renditions) at three different tempi, averaged across the three performances within each tempo category.**

Nevertheless, further statistical analysis revealed that there were some reliable changes in temporal profiles across tempi. Analyses of variance were first conducted on each pianists' complete log-transformed IOI data, with IOI (214 levels; see footnote 4) and tempo (3 levels) as fixed factors, and individual performances nested within tempi (3 levels) as the random factor. If the log-transformed timing profiles are strictly parallel, then the IOI by tempo interaction should be nonsignificant. In other words, the between-tempo variation in profile shape should not significantly exceed the within-tempo variation. For LPH, this was indeed the case [$F(426,1278) = 1.03$]. For BHR, however, there was a small but significant interaction [$F(426,1278) = 1.55$, $p < .0001$]. Moreover, when similar analyses were conducted on bars 0-8, corresponding to the excerpts used in the perceptual test (though with all three performances at each tempo included), the IOI by tempo interaction was significant for both LPH [$F(96,288) = 1.57$, $p < .003$] and BHR [$F(96,288) = 1.77$, $p < .0003$]. Additional analyses confirmed that BHR showed subtle changes in timing pattern with tempo throughout most of the music, whereas LPH showed no statistically reliable changes with tempo except at the very beginning of the piece.[6]

Yet another way of analyzing the data revealed, however, that LPH did show some systematic deviations from relational invariance, after all. For pairs of tempi, the log-transformed average IOIs at the faster tempo were subtracted from the corresponding IOIs at the slower tempo, and the correlation between these differences and the IOIs at the slower tempo was examined. The relational invariance hypothesis predicts that the log-difference (i.e., the ratio) between corresponding IOIs should be constant, and hence the correlation with IOI magnitude should be zero. However, LPH showed a highly significant positive correlation for medium versus fast tempo [$r(252) = 0.50$, $p < .0001$], which indicates that long IOIs changed disproportionately more than short IOIs. BHR showed a similar but smaller tendency, which was nevertheless significant [$r(252) = 0.18$, $p < .01$]. Between the slow and medium tempi, LPH showed a weak tendency in the opposite direction, whereas BHR showed no significant correlation. Very similar results were obtained for bars 1-8 only; thus the correlations did not derive solely from the very long IOIs. Relational invariance evidently held to a greater degree between the slow and medium tempi than between the medium and fast tempi.

Thus there is statistical evidence that relational invariance did not hold strictly, as observed also by Heuer (1991) in his discussion of simpler motor tasks. Still, the timing patterns were highly similar across the different tempi, and the differences that did occur do not suggest different structural interpretations. Moreover, they were not sufficiently salient perceptually to enable listeners to discriminate original performances from performances whose tempo had been modified artificially. (Since there was no evidence that changes with tempo were larger later in the piece than at the beginning, the results of the perceptual test for bars 0-8 probably can be generalized to the whole performance.) It seems fair to conclude, then, that relational invariance of timing profiles held *approximately* in these performances.

Grace note timing patterns

The analysis so far has focused on eighth-note (and longer) IOIs only. However, "Träumerei" also contains grace notes. The question of interest was whether the timing of the corresponding tones also changed proportionally with overall tempo.

There are two types of grace notes in the piece. Those of the first type, notated as small eighth-notes, are important melody notes occurring during major ritardandi. The slowdown in tempo lets them "slip in" without disturbing the rhythm. One of these notes occurs in bar 8 (between the fourth and fifth eighth-notes); the other is the upbeat to the recapitulation in bar 17 (following the last eighth-note in bar 16). The other type, notated as pairs of small sixteenth-notes, represents written-out left-hand arpeggi (bars 2, 6, and 18) and is played correspondingly faster. While proportional scaling of the first, slow type was to be expected because they are essentially part of the expressive timing profile, the prediction for the second type was less clear. However, neither type is readily classifiable in terms of Desain and Honing's (1992b) appoggiatura-acciaccatura distinction.

Several ANOVAs were conducted on the log-transformed IOIs defined by these grace notes. Separate analyses of bars 8 and 16 included three IOIs: the preceding eighth-note IOI and the two parts of the eighth-note IOI bisected by the onset of the grace note. In neither case was there a significant IOI by tempo interaction.[7] Thus the timing of these grace notes scaled proportionally with changes in tempo. LPH and BHR differed in

their timing patterns: In bar 8, the grace note occupied 68% (LPH) versus 57% (BHR) of the fourth eighth-note IOI; in bar 16, it occupied 55% (LPH) versus 37% (BHR) of the last eighth-note IOI. BHR thus took the grace-note notation somewhat more literally than did LPH.

Bars 2 and 18 were analyzed separately from bar 6, which involved different pitches (cf. Figure 1), though the timing patterns were found to be very similar. In each case there were three IOIs, defined by the onset of the second eighth-note in the right hand, the onsets of the two grace notes in the left hand, and the onset of the highest note of the chord in the right hand. In no case was there a significant IOI by tempo interaction.[8] This, even for these *arpeggio* grace notes relational invariance seemed to hold. There were again individual differences between the two pianists: In bars 2, 6, and 18, the two grace notes together took up 68% (LPH) versus 79% (BHR) of the second eighth-note IOI, while the second grace note occupied 63% (LPH) versus 69% (BHR) of the grace note interval. Both pianists maintained their individual patterns across bars 2, 6, and 18. Both individual timing patterns are within the range of typical values observed in Repp's (1992) analyses of famous pianists' performances.[9]

Chord asynchronies

One of the factors that might enable listeners to discriminate a slowed-down fast performance from an originally slow one is that the asynchronies among nominally simultaneous tones are enlarged in the former. It was surprising, therefore, that there was no indication in the perceptual data that slowed-down fast performances sounded worse than speeded-up slow ones. There is no obvious reason why unintended vertical asynchronies (i.e., purely technical inaccuracies) should scale with tempo. However, asynchronies that are intended for expressive purposes (cf. Vernon, 1937; Palmer, 1989) might possibly increase as tempo decreases. Rasch (1988) reports such a tendency for asynchronies in ensemble playing.

A complete analysis of asynchronies was beyond the scope of this paper. Instead, the analysis focused on selected instances only. One good candidate was the four-tone chord near the beginning of each phrase (see Figure 1). This chord occurs eight times during the piece, seven times in identical form in F major and once (bar 13) transposed up a fourth to B-flat major. It does not contain any melody tones, which eliminates one major motivation for expressive asynchrony (cf. Palmer, 1989). However, it is technically tricky because it involves both hands in interleaved fashion: The first and third tones (numbering from top to bottom) are taken by the right hand, while the second and fourth tones are taken by the left hand. In addition to unintended asynchronies that may be caused by this physical constellation, planned asynchronies may be involved in the proper "voicing" of the chord, which is a sonority of expressive importance.

For each of the eight instances of the chord in each performance, the IOIs for the three lower tones (Tones 2-4) were calculated relative to the highest tone (Tone 1).[10] Lag times were positive whereas lead times were negative. The untransformed IOIs were subjected to separate ANOVAs for each pianist, with the factors rendition (i.e., the 8 occurrences of the chord), tone (3), and tempo (3); performances (3, nested within tempo) were the random variable. The question here was whether the IOIs changed at all with tempo.

The ANOVA results were similar for the two pianists. Both LPH [$F(1,6) = 70.94$, $p < .0003$] and BHR [$F(1,6) = 84.71$, $p < .0002$] showed a highly significant grand mean effect, indicating that the average IOI was different from zero: The three lower tones lagged behind the highest tone by an average of 16.9 ms (LPH) and 9.6 ms (BHR), respectively. In addition, each pianist showed a rendition main effect [LPH: $F(7,42) = 3.58$, $p < .005$; BHR: $F(7,42) = 12.40$, $p < .0001$], a tone main effect [LPH: $F(2,12) = 12.80$, $p < .002$; BHR: $F(2,12) = 49.18$, $p < .0001$], and a rendition by tone interaction [LPH: $F(14,84) = 2.56$, $p < .005$; BHR: $F(14,84) = 4.34$, $p < .0001$]. However, no effects involving tempo were significant. In particular, there was no indication that lags were larger at the slow tempo. The high significance levels of the other effects show that the data were not simply too variable for systematic effects of tempo to be found.

The systematic differences in patterns of asynchrony across renditions are of theoretical interest and warrant a brief discussion. They are shown in Figure 4, averaged across all performances. The mean lags for individual tones indicated that, for both pianists, the lowest (fourth) tone lagged behind more than the other two. For BHR, moreover, the third tone did not lag behind at all, on the average. This was clearly a reflection of the fact that it was played with the same hand as the highest (first) tone.

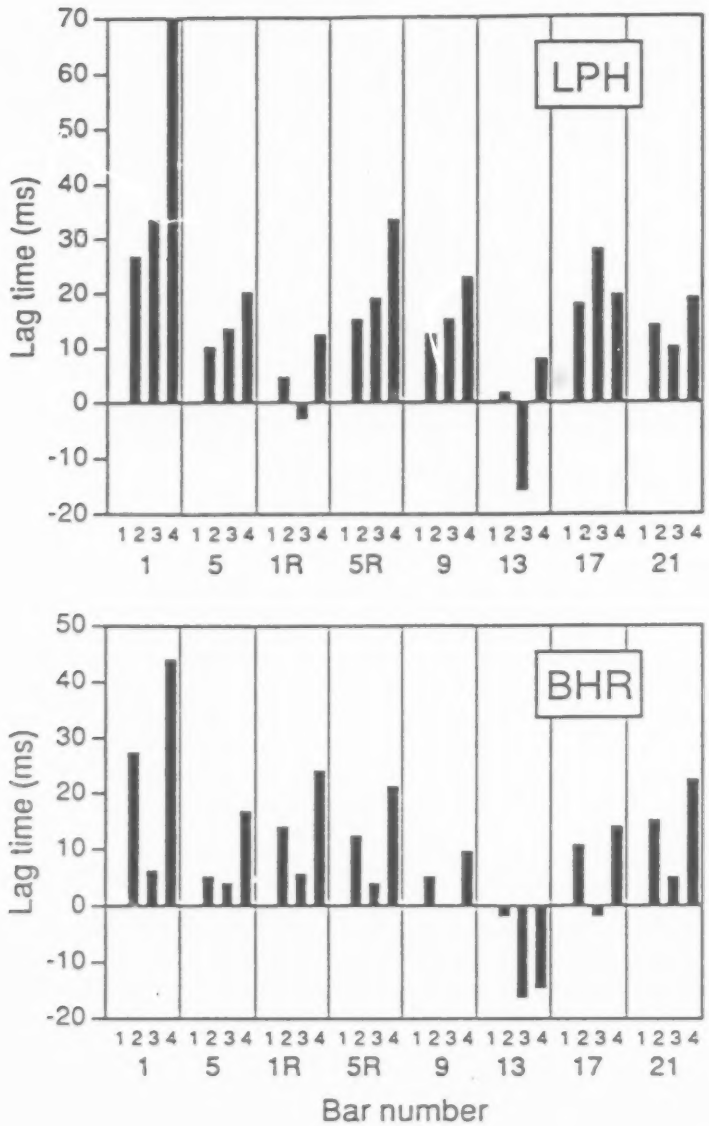


Figure 4. **Chord asynchronies relative to the highest tone, averaged across all tempi. The bars for each chord show lag times for tones 2-4 (in order of decreasing pitch) relative to tone 1. "R" after a bar number stands for "repeat."**

Both pianists showed the largest lag times in bar 1, which is perhaps a start-up effect of "getting the feel" of the instrument. Again for both pianists, the only rendition in which any lower tones *preceded* the highest tone was the one in bar 13, which is in a different key. The precedence of the third tone here may be due to the fact that it falls on a black key, which is more elevated than the white keys and hence is reached earlier by the thumb of the right hand. Apart from this exceptional instance, BHR maintained his characteristic 1-3-2-4 pattern across all renditions, though the magnitude of the lags varied somewhat. LPH was more variable, showing an average 1-2-3-4 pattern in four renditions, a 1-3-2-4 pattern in two, and a 1-2-4-3 pattern in one.

It could well be that the chord asynchronies just analyzed did not vary with tempo because they were not governed by expressive intent, only by fingering and hand position. The chord examined did not contain any melody tones and thus did not motivate any highlighting of voices through deliberate timing offsets. Therefore, a second set of chords was examined. The tone clusters on the second and third eighth-notes of bars 10 and 14 (see Figure 1) also contain four tones each (three in one instance), but all of them have melodic function. The soprano voice completes the primary melodic gesture; the alto voice "splits off" into a counterpoint; the tenor voice enters with an imitation of the primary melody, accompanied by the bass voice, which has mainly harmonic function. Thus the relative salience of the voices may be estimated as 1-3-2-4, which may well be reflected in a sensitive pianist's relative timing of the tone onsets. Unlike the chords examined previously, tones 1 and 2 are played by the right hand, whereas tones 3 and 4 are played by the left hand. Since the highlighting of the tenor voice entry (left hand) may be the performer's primary goal, the temporal relationship between the two hands may be the primary variable here.

This indeed turned out to be the case, though with striking differences between the two pianists. LPH changed her pattern of asynchronies radically (and somewhat inconsistently) with tempo in bar 10; otherwise, however, the patterns did not vary significantly with tempo. The main effects of the tone factor, on the other hand, were generally significant, indicating consistency in patterning for each chord.

The results are shown in Figure 5. As can be seen, LPH showed enormous asynchronies which were almost certainly intended for expressive purposes. They alternated between two patterns: The left hand (tones 3 and 4) either lagged behind or led the right hand (tones 1 and 2). The right-before-left pattern occurred in positions 10-2 (slow tempo) and 14-2. The left-before-right pattern occurred in positions 10-3 (at both slow and fast tempo) and 14-3, as well as in position 10-2 at the fast tempo. The medium-tempo patterns in positions 10-2 and 10-3 (not shown in the figure) were split among the two alternatives. The two tones in each hand were roughly simultaneous, except in position 14-3, where the lowest note occurred especially early. The net effect in bar 14 and, less consistently, in bar 10 was that the tenor and bass voices entered late but continued early, so that the first IOI in these voices was considerably shorter than the corresponding IOI in the soprano voice. This makes good sense: While the soprano voice slows down towards the end of a gesture, the tenor voice initiates the imitation with a different temporal regime. The discrepancy thus highlights the independent melodic strains.

The statistical analysis was conducted on the left-hand lag times only, because they showed the major variation. Due to LPH's strategic changes with tempo in bar 10, there were highly significant interactions with tempo in the overall analysis of her data. A separate analysis on bar 14, however, revealed no effects involving tempo. The difference between positions 14-2 and 14-3 was obviously significant, as was the effect of tone [$F(1,6) = 19.57$, $p < .005$] and the position by tone interaction [$F(1,6) = 28.34$, $p < .002$].

BHR, on the other hand, showed a very different pattern. His asynchronies were much smaller and always showed a slight lag of the left hand, more so in bar 10 than in bar 14. His patterns were highly consistent but reveal little expressive intent, as they are comparable in magnitude to the asynchronies in the nonmelodic chord analyzed earlier. The statistical analysis on the left-hand lag times showed the difference between bars 10 and 14 to be highly reliable [$F(1,6) = 24.51$, $p < .003$]. The tendency for the bass voice to lag behind the tenor voice in bar 10, but to lead it in bar 14, was reflected in a significant interaction [$F(1,6) = 12.66$, $p < .02$]. No effect involving tempo was significant, however. Thus, neither pianist's patterns scaled with tempo. Relational invariance does not seem to hold at this level; instead, there is absolute invariance across tempi, except for the qualitative changes noted in some of LPH's data.

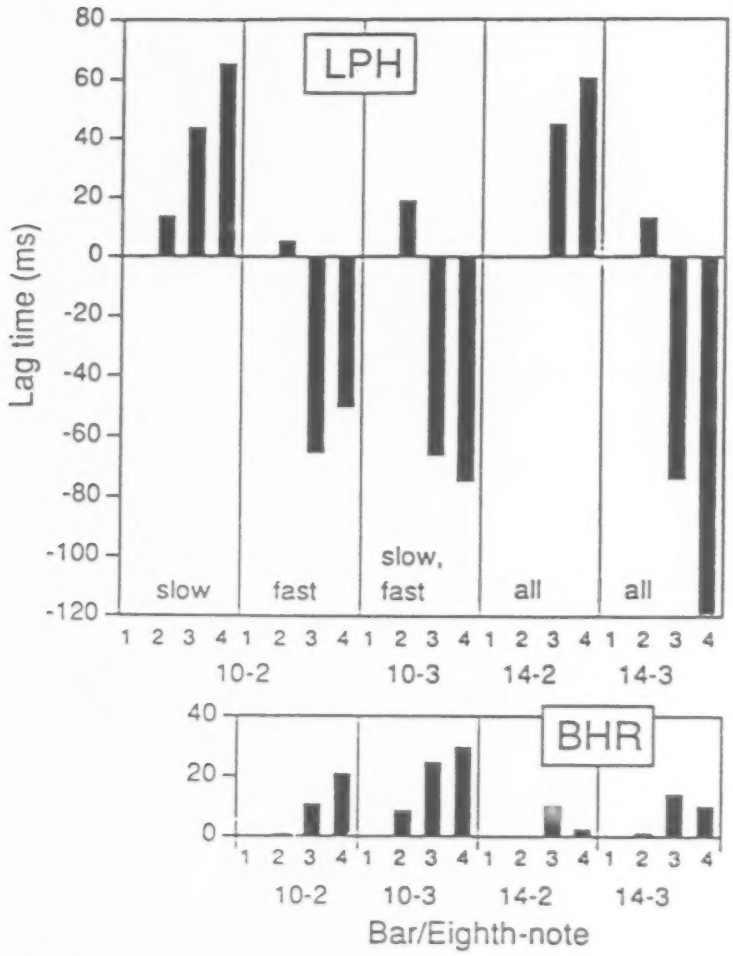


Figure 5. **Chord asynchronies in bars 10 and 14, averaged across all tempi except as noted. The second tone is absent in position 14-2.**

Tone overlap

Another detailed feature that was examined, again in a selective manner, was tone overlap. MacKenzie and Van Eerd (1990), in their study of piano scale playing, found that the absolute overlap of successive tones in legato scales decreased as tempo increased. At the slowest tempo (nominally similar to the present slow tempi, but really twice as fast because the scales moved along in sixteenth-notes) successive tones overlapped by about 15 ms, which decreased to near zero as tempo increased. Because of the expressive legato required in "Träumerei," and the slower rate of tones (eighth-notes), larger overlaps were to be expected. The question was whether they got smaller as tempo increased. The measure of overlap was the difference between the offset time of one tone and the onset time of the following tone; a negative value indicates a gap between the tones (which may not be audible because of pedalling).

The passage selected for analysis was the ascending eighth-note melody in the soprano voice which occurs first in bars 1/2 and recurs in identical or similar form 7 times during the piece (see Figure 1). Because of different fingering and an additional voice in the right hand, the instances in bars 9/10 and 13/14 were not included. Moreover, because overlap times varied wildly in the last interval of bar 22 (preceding the fermata) for both pianists, the data were restricted to the first four instances only (i.e., bars 1/2 and 5/6, and their repeats). Each of these consists of 5 consecutive eighth-notes, yielding 4 overlap times. The last two notes span the interval of a fourth in two instances (bars 1/2) and a major sixth in the other two (bars 5/6); otherwise, the notes are identical. The fingering is likely to be the same. The statistical analyses (on untransformed data) included bars as a factor, as well as renditions (here, strict repeats), tones, and tempo. Again, the question was whether the overlap times varied with tempo at all.

LPH's data were quite a surprise because her overlap times were an order of magnitude larger than expected. Not only did she play legatissimo throughout, but she often held on to the second and fourth tones through the whole duration of the following tone, resulting in overlap times in excess of 1 s. Since these overlap times thus became linked to the regular eighth-note IOIs, it is not surprising that effects of tempo were found. The effects were not simple, however. The ANOVA showed all main effects and interactions involving bar, tone, and tempo (except for the bar by tone interaction) to be significant. The data corresponding to the triple interaction [$F(6,18) = 4.04$, $p < .01$] are shown in Figure 6 below. The data are averaged over renditions, which had only a main effect [$F(1,6) = 10.30$, $p < .02$], representing an increase in overlap with repetition.

Perhaps the most interesting fact about LPH's data is that the overlap pattern differed between bars 1/2 and 5/6 from the very beginning, even though the two tone sequences were identical up to last tone. There was slightly more overlap between the first two tones in bars 1/2 than in bars 5/6. While this may not be a significant difference, there is no question that the overlap between the following two tones was much greater in bars 1/2 than in bars 5/6. The same is true for the next two tones, although there is much less overlap in absolute terms. Finally, the situation is reversed for the last two tones, which show enormous overlap in bars 5/6, but much less in bars 1/2. The tempo effects within this striking interaction are fairly consistent (overlap decreasing with increasing tempo), except for the third and fourth tones in bars 1/2, whose overlap *increases* with tempo—an unexplained anomaly.

BHR's overlap times, in contrast to LPH's, were of the expected magnitude, and there was no significant effect involving tempo. However, the main effects of bar and tone, and the bar by tone interaction [$F(3,18) = 24.90$, $p < .0001$] were all highly significant. As for LPH, the difference in final interval size affected the overlap pattern of the whole melodic gesture. BHR showed negative overlap (i.e., a gap, camouflaged by pedalling) between the first two tones, which was larger in bars 5/6 than in bars 1/2. The overlap between the second and third tones was larger in bars 5/6, but that between the third and fourth tones was larger in bars 1/2. Finally, there was overlap between the last two tones in bars 1/2, but a small gap in bars 5/6, probably reflecting the stretch to the larger interval of a sixth. BHR's playing style thus can be seen to be only imperfectly *legato*, perhaps due to his less developed technique. Nevertheless, he was highly consistent in his imperfections, and his data suggest independence of tempo.

Pedal timing

As the last temporal facet of these performances, we consider the timing of the sustain pedal. Although variations in pedal timing within tone IOIs probably had little if any effect on the acoustic output, they are of interest from the perspective of motor organization and control.

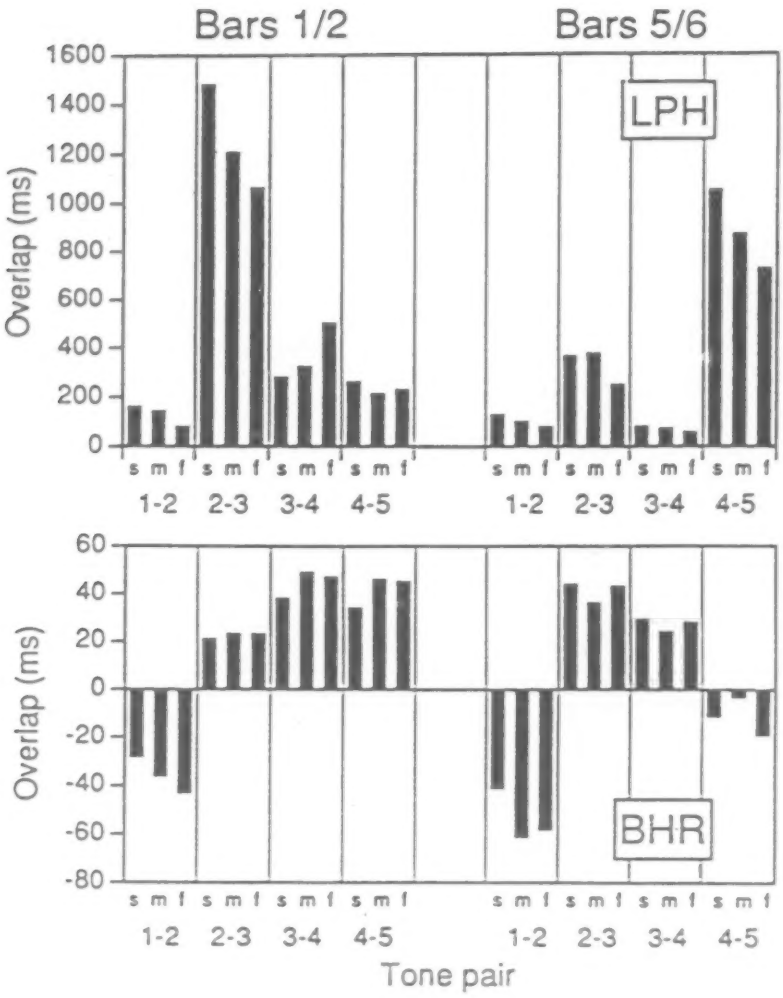


Figure 6. **Tone overlap times in bars 1/2 and 5/6 at s(low), m(edium), and f(ast) tempo.**

The pianist must coordinate the foot movements with the hand movements, and this is usually done subconsciously, while attention is directed towards the keyboard.

A complete analysis of the pedalling patterns would lead too far here, as the pedal was used continuously throughout each performance.[11] The analysis focused on the opening motive of each phrase, specifically on the interval between the downbeat melody tone and the following 4-note chord, whose highest tone was used as the temporal reference (see Figure 1). This quarter-note IOI, which occurred eight times in the course of each performance (in bars 1, 5, 1R, 5R, 9, 13, 17, and 21, where "R" stands for "repeat"), usually contained a pedal change (i.e., a pedal offset followed by a pedal onset) in both pianists' performances, though BHR's pedal offsets generally fell very close to the onset of the downbeat tone, often preceding it slightly. Exceptions were bars 1, 1R, and 9, which usually contained only a pedal onset, the previous offset being far outside the IOI or absent (at the beginning of the piece). Only the pedal onsets are marked in the score. The pedal serves here to sustain the bass note struck on the downbeat, which a small left hand needs to relinquish in order to reach its portion of the following chord. In bar 13, in fact, the bass note can be sustained only by pedalling, as it occurs in a lower octave.

The question of interest was whether, across the variations in tempo, the pedal actions occurred at a fixed time after the onset of the quarter-note IOI or whether their timing varied in proportion with tempo, so that they occurred at a fixed percentage of the time interval. Figure 7 shows the average quarter-note IOI durations (squares) at the three tempi in the different bars. As expected, there was generally a systematic decrease in IOI duration from slow to medium to fast tempo [LPH: $F(2,6) = 12.85$, $p < .007$; BHR: $F(2,6) = 88.25$, $p < .0001$]. Both pianists also showed a highly significant main effect of bar on IOI duration [LPH: $F(7,42) = 21.36$, $p < .0001$; BHR: $F(7,42) = 47.38$, $p < .0001$]. The longest IOIs occurred in bar 17 (the beginning of the recapitulation); LPH also had relatively long IOIs in bar 21 and short ones in bar 1, whereas BHR had relatively long IOIs in bar 13 (the passage in a different key). There was a small tempo by bar interaction for BHR only [$F(14,42) = 2.14$, $p < .03$].

Figure 7 also shows average pedal offset and onset times, as well as bass tone offset times. The two pianists differed substantially in their pedalling strategies. LPH's pedal offsets (when present) occurred well after IOI onset, and her pedal onsets occurred around the middle of the IOI. BHR's pedal offsets, on the other hand, occurred at the beginning of the IOI and his pedal onsets were relatively early also. Another striking difference between the two pianists is that LPH held on to the bass tone (as indicated by the absence of dashes in the figure), except in bar 13 where this was physically impossible. BHR, on the other hand, despite his larger hands, always released the bass tone following the pedal depression.

For LPH, pedalling did not to seem to vary systematically with tempo.[12] This was confirmed in ANOVAs on pedal offset and onset times, as well as on their difference (pedal off time). None of these three variables showed a significant tempo effect or a tempo by bar interaction. However, all three varied significantly across bars [$F(4,24) = 8.57$, $p < .0003$; $F(7,42) = 15.43$, $p < .0001$; $F(4,24) = 10.00$, $p < .0002$]. For BHR, on the other hand, both pedal offset times [$F(2,6) = 10.06$, $p < .02$] and pedal onset times [$F(2,6) = 6.57$, $p < .04$] decreased as tempo increased, whereas pedal off times did not vary significantly. All three varied significantly across bars [$F(4,24) = 6.09$, $p < .002$; $F(6,36) = 12.58$, $p < .0001$; $F(4,24) = 22.03$, $p < .0001$], even though the differences were much smaller than for LPH. Even more clearly than the pedalling times, BHR's bass tone offset times varied with tempo [$F(2,6) = 27.26$, $p < .001$], and also across bars [$F(7,42) = 14.14$, $p < .0001$].

To determine whether BHR's timings exhibited relational invariance, they were expressed as percentages of the total IOI and the ANOVAs were repeated. Of course, the pedal offset times, which varied around the IOI onset, could not be effectively relativized in this way, so that the tempo effect persisted [$F(2,6) = 9.36$, $p < .02$]. The tempo effect on pedal onset times, on the other hand, disappeared [$F(2,6) = 0.37$], whereas a tempo effect on pedal off times emerged [$F(2,6) = 12.62$, $p < .008$]. Thus BHR's data could be interpreted as either exhibiting relational invariance for pedal onset times, or absolute invariance of pedal off times. There was no question, however, that BHR's bass tone offset times scaled proportionally with tempo, as there was no tempo effect on the percentage scores [$F(2,6) = 0.71$]. All relativized measures, on the other hand, continued to vary significantly across bars [$F(4,24) = 4.62$, $p < .007$; $F(6,36) = 17.81$, $p < .0001$; $F(4,24) = 23.86$, $p < .0001$; $F(7,42) = 18.90$, $p < .0001$]. This was equally true for LPH's data.

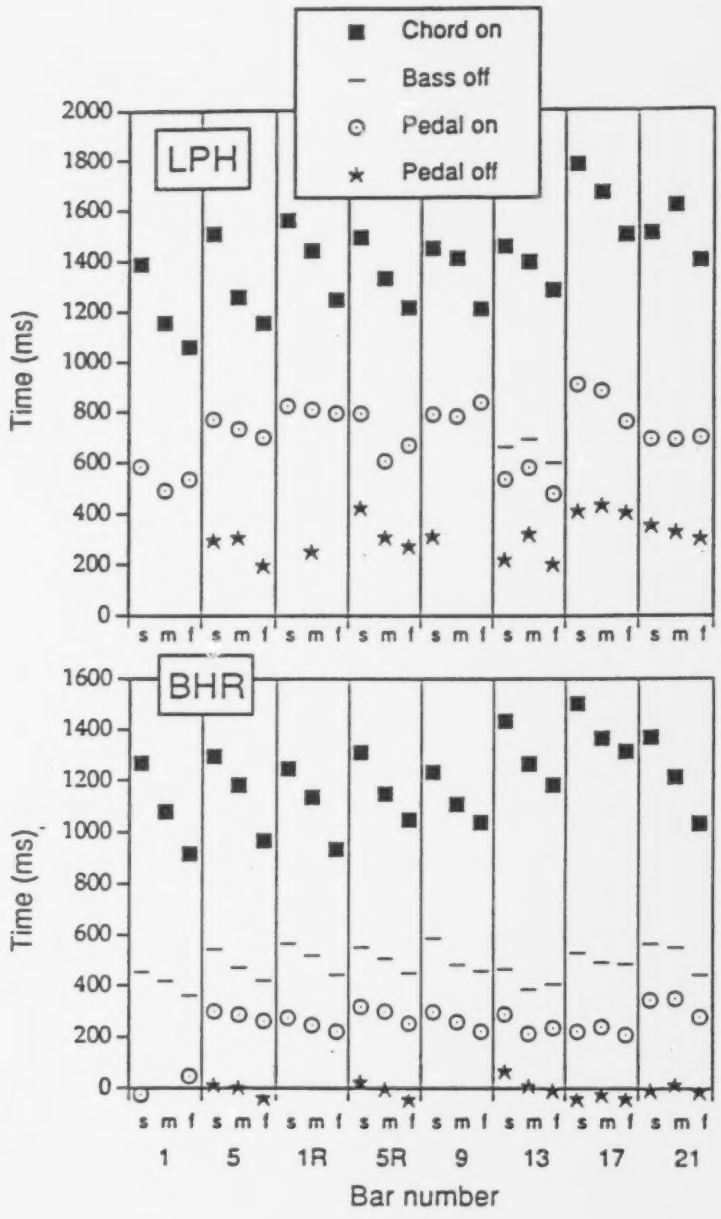


Figure 7. **Absolute pedal offset and onset times and bass tone offset times within the first (quarter-note) IOI in each of eight bars, at s(low), m(edium), and f(ast) tempo. Each IOI starts at 0 and ends with the onset of (the highest tone of) the chord. Missing symbols reflect incomplete or absent data.**

In summary, these data do not offer much support for the hypothesis of relational invariance in pedal timing across tempo changes. LPH's data were highly variable. BHR's data are consistent with an alternative hypothesis—that the pedal off-on action (a rapid up-down movement with the foot) was independent of tempo but was initiated earlier at the faster tempo. This earlier start may be a reflection of a general increase in muscular tension.

The results are much clearer with regard to the differences in pedalling times across bars. Clearly, these effects are *not* relationally invariant and thus must be caused by structural or expressive factors that vary from bar to bar. This may be the first time that such systematic variation has been demonstrated in pedalling behavior—an interesting subject for future studies.

Intensity microstructure

A detailed analysis of the intensity microstructure of the performances would lead too far in the present context. To verify, however, that both pianists showed meaningful patterns of intensity variation, Figure 8 shows the intensity profiles (strictly speaking, MIDI velocity profiles) of the melody notes (soprano voice) averaged across the three medium-tempo performances of each pianist (and averaged across the two renditions of bars 1-8). The format is similar to that of Figure 2, with corresponding phrases superimposed.

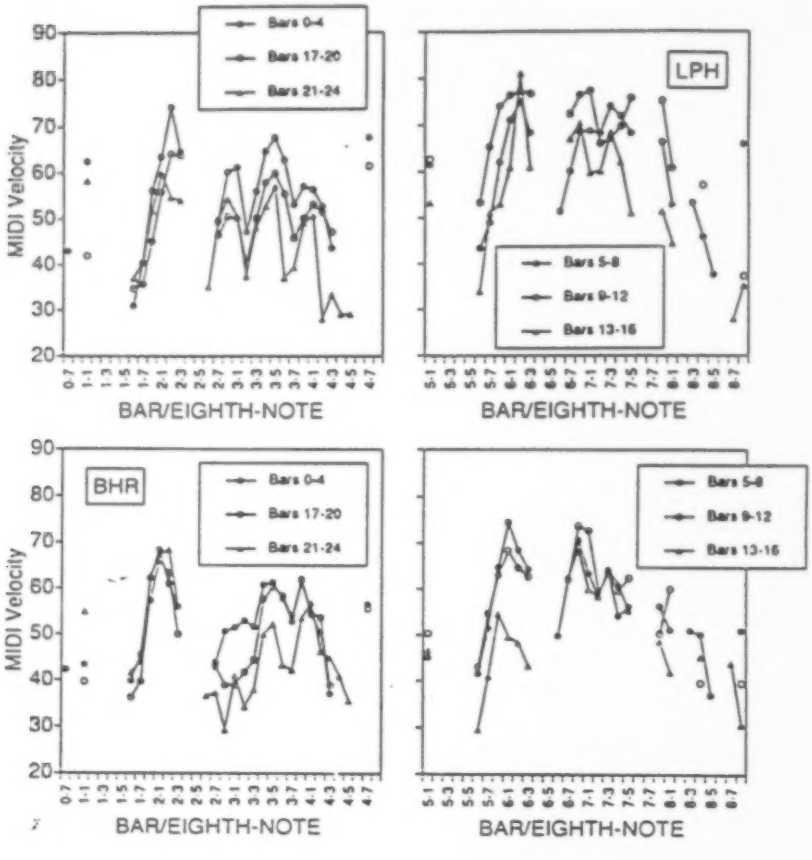


Figure 8. **Average MIDI velocity as a function of metric distance, averaged across three medium-tempo performances. Structurally identical or similar phrases are overlaid. Only adjacent eighth-notes are connected.**

The intensity patterns were slightly more variable than the timing patterns. Nevertheless, both pianists showed structurally meaningful patterns that were similar across phrases of similar structure.[13] For example, the salient melodic ascent in bars 1-2, 5-6, and so on, always showed a steep rise in intensity which leveled off or fell somewhat as the peak was reached. The motivic chain during the second part of each phrase showed modulations related to the motivic structure, especially in the phrase type shown in the left-hand panels, and a steady *decrescendo* in the phrase type shown in the right-hand panels. Both pianists played bars 17-20 somewhat more softly than the identical bars 0-4, and ended the final phrase (bars 21-24) even more softly (left-hand panels). LPH, but not BHR, played bars 9-12 more strongly than bars 5-8 (right-hand panels). Both pianists played bars 13-16 more softly than bars 9-12; BHR especially observed the *pp* marking in bar 12 in the Clara Schumann edition (cf. Figure 1).

Given these fairly orderly patterns, the question of interest was: Did they vary systematically with tempo? Before addressing this issue, however, it may be asked whether the average dynamic level was equal across tempi. According to Todd (1992), a faster tempo may be coupled with a higher intensity.

Figure 9 shows that this was indeed the case in the present performances: For both LPH and BHR, the average intensity of the soprano voice increased from slow to fast and from medium to fast tempo; there is no clear difference between slow and medium for LPH. The magnitude of the change is not large—about 2 MIDI units, which translates into about 0.5 dB (Repp, 1993). However, there is a second systematic trend in the data for both pianists: In the course of the recording session, the average intensity dropped by an amount equal to (BHR) or larger (LPH) than that connected with a tempo change. Interestingly, there was also a tendency for the tempi to slow down across repeated performances (see Repp, in press), but that decrease was much smaller than the difference among the principal tempo categories. Both trends may reflect the pianists' progressive relaxation and adaptation to the instrument; possibly also fatigue. Finally, it is clear from Figure 9 that BHR played more softly, on the average, than LPH.[14]

Because of the within-tempo variation in average intensity, the between-tempo differences were nonsignificant, though they appeared to be systematic. For the same reason, the statistical assessment of variations in intensity patterns across tempi was somewhat problematic, for if that variation was contingent on overall intensity ("level of tension") there was no reason to expect it to be larger between than within tempi. ANOVAs were nevertheless carried out. Across all melody tones (N=166), the tone by tempo interaction fell short of significance for LPH [$F(330,990) = 1.13$, $p < .08$] and was clearly nonsignificant for BHR. Across the melody tones of the excerpt used in the perceptual test (bars 0-8, N=42), the tone by tempo interaction reached significance for LPH [$F(82,246) = 1.49$, $p < .01$] but remained clearly nonsignificant for BHR.[15]

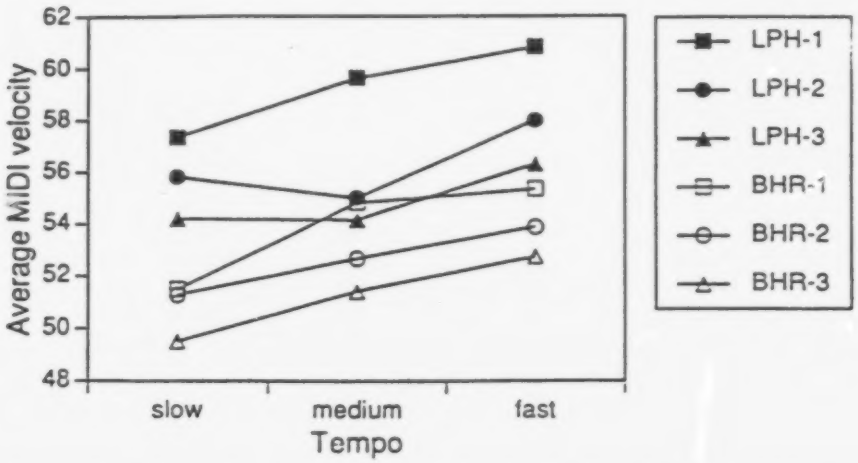


Figure 9. **Grand average MIDI velocity as a function of tempo, separately for each pianist's three performances at each tempo.**

Figure 10 shows the average intensity profiles for the melody tones in bars 0-8 at the three tempi for each pianist. The similarities among the three profiles are far more striking than the differences. For LPH, one source of the significant tone by tempo interaction is evident at the very beginning: The dynamic change from the upbeat to the downbeat was far larger at the fast tempo than at the slow tempo. (Curiously, BHR showed the opposite.) On the whole, however, the intensity profiles remained invariant across tempi.

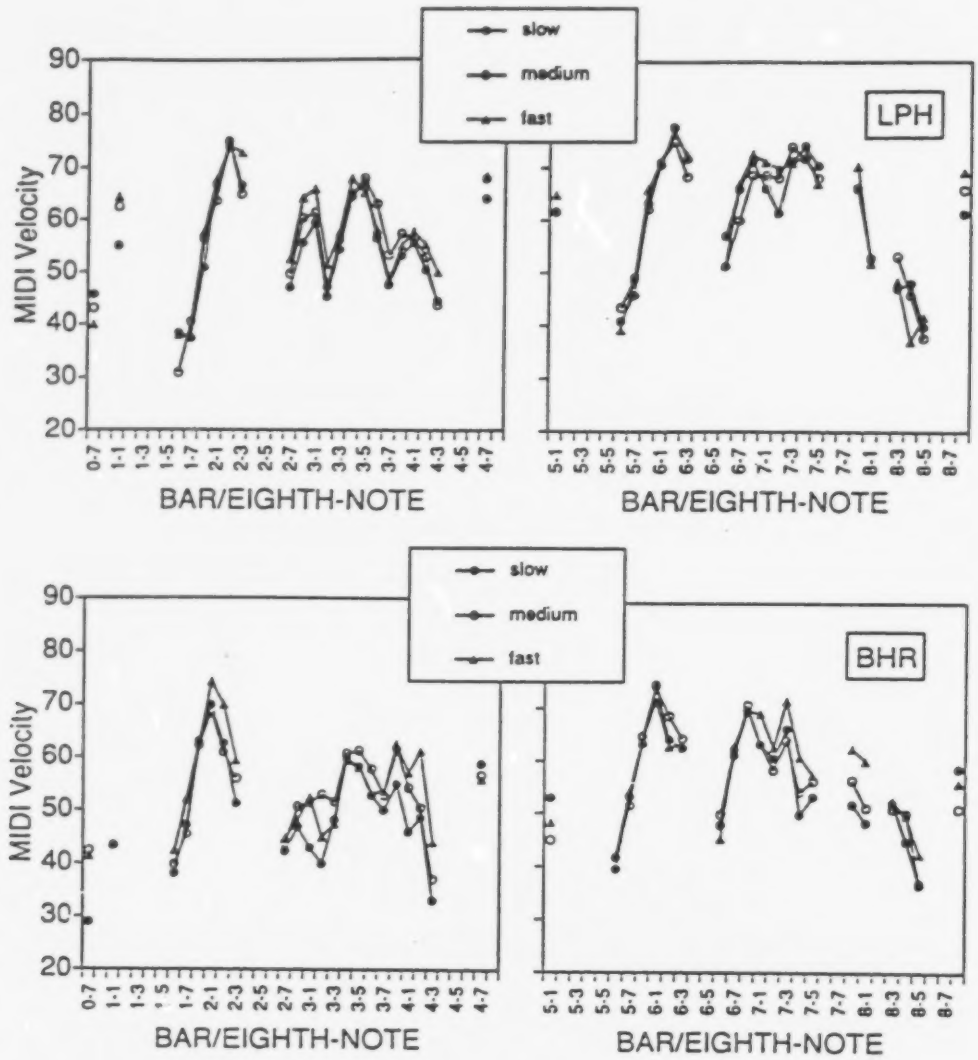


Figure 10. **Average MIDI velocity profiles for bars 0-8 at three different tempi.**

GENERAL DISCUSSION

The present study explored in a preliminary way the question of whether the expressive microstructure of a music performance remains relationally invariant across moderate changes in tempo. The results suggest that the onset timing and intensity profiles essentially retain their shapes (the latter being very nearly constant), whereas other temporal features are largely independent of tempo and hence exhibit local constancy rather than global relational invariance.

After a perceptual demonstration that a uniform tempo transformation does no striking damage to the expressive quality of performances of Schumann's "Träumerei," six specific aspects of the expressive microstructure were examined. The first and most important one was the pattern of tone onset times. That pattern was highly similar across tempi, yet deviated significantly from proportionality. Although relational invariance did not hold perfectly, the deviations seemed small and not readily interpretable in terms of structural reorganization. The second aspect examined was the timing of grace notes. Somewhat surprisingly, in view of the observations by Desain and Honing (1992a, 1992b) that stimulated the present study, grace notes were timed in a relationally invariant manner across tempi. This may have been due to their relatively slow tempo and expressive function; the result may not generalize to ornaments that are executed more rapidly. The third aspect concerned selected instances of onset asynchrony among the tones of chords. These asynchronies did not change significantly with tempo, even when they were rather large and served an expressive function; apparently, they were "anchored" to a reference tone. The fourth aspect was the overlap among successive *legato* tones. When this overlap was of the expected magnitude (pianist BHR), it was independent of tempo. For one pianist (LPH), the overlaps were unusually large and did vary with tempo. The fifth aspect examined was pedal timing. It did not seem to depend on tempo in a very systematic way and had no audible effect in any case. Finally, the intensity microstructure was studied. Although overall intensity increased somewhat with tempo, the detailed intensity pattern remained nearly invariant.

In no case were there striking deviations from relational invariance. Those features that were found not to scale proportionally with tempo were generally on a smaller time scale, so that artificially introduced proportionality (in the perceptual test) was difficult to detect. Because of the limited contribution of these details to the overall impression of the performances, it may be concluded that relational invariance of expressive microstructure held approximately across the tempi investigated here. This is consistent with the notion that musical performance is governed by a "generalized motor program" (Schmidt, 1975) or "interpretation" that includes a variable rate parameter. Although there are good reasons for expecting the interpretation to change when tempo changes are large (Clarke, 1985; Handel, 1986), the tempo range investigated here apparently could accommodate a single structural interpretation.

All microstructural patterns examined exhibited large and highly systematic variation as a function of musical structure, as well as clear individual differences between the two pianists. While the data on timing microstructure (including grace notes) are consistent with the detailed analyses in Repp (1992), the observations on chord asynchrony, tone overlap, pedalling, and intensity microstructure go well beyond these earlier analyses. However, since the focus of the present study was on effects of tempo, the structural interpretation of microstructural variation was not pursued in great detail here. Suffice it to note once again the pervasive influence of musical structure on performance parameters and the astounding variety of individual implementations of what appear to be qualitatively similar structural representations.

Despite certain methodological parallels, the present study is diametrically opposed to the scale-playing experiment of MacKenzie and Van Eerd (1990), which essentially focused on pianists' inability to achieve a technical goal (viz., mechanical exactness). By contrast, the primary concern here was pianists' success in realizing the artistic goal of expressive performance. The relative invariances observed in timing and intensity patterns were cognitively governed, not due to kinematic constraints. However, both types of studies have their justification and provide complementary information. At some of the detailed levels examined here (chord asynchronies, tone overlap) motoric constraints such as fingering patterns clearly had an influence. Thus, even within a slow, highly expressive performance, there are aspects that are governed to some extent by "raw technique," and ultimately the technical skills that help an artist to achieve subtlety of expression may be the very same that help her/him to play scales smoothly.

However, the technique serves its true purpose only in the context of an expressive performance and therefore is perhaps better studied in that context also.

The present study has limitations, the most obvious of which is that only a single piece of music was examined. Clearly, the character of "Träumerei" is radically different from that of Salieri's "Nel cor piú non mi sento," whose timing microstructure Desain and Honing (1992a) found to change considerably with tempo. "Träumerei" is slow, expressive, entirely *legato*, rhythmically vague, and contains few fast notes; even the grace notes are timed deliberately. The Salieri ditty, on the other hand, is moderately fast, requires a detached articulation and rhythmical precision, and contains a number of very fast grace notes (*acciaccature*) that may resist further reduction in duration. Clearly, it would be wrong to conclude from the present findings that expressive microstructure *always* remains invariant across tempo changes, even to a first approximation. Desain and Honing's results, in spite of their informality, already seem to provide a counterexample.[16] So do perhaps Clarke's (1982, 1985) findings. The conclusion seems justified, however, that in *some* types of music relational invariance is maintained to a large extent. This music is probably the kind that carries expressive gestures on an even rhythmic flow. Rhythmically differentiated music containing many short note values, rests, and varied articulation (such as examined by Desain and Honing, 1992a) is probably much more susceptible to changes with tempo.

In fact, it would be both more precise and more cautious to say that relational invariance *can* be maintained in certain kinds of music. The present pianists, although they did not try very consciously to maintain their "interpretation" across tempi but instead attempted to play as naturally and expressively as possible, nevertheless refrained from "changing their interpretation" in the course of the recording session, whatever that may imply in an objective sense. In real-life music performance, on the other hand, changes in tempo are probably more often a consequence of a change in an artist's conception of the music than a deliberate change in the tempo as such. Tempo differences observed among different artists are often to be understood in this way. What the present findings imply, however, is that if different artists were forced to play at the same overall tempo, without giving them time to reconsider their "interpretation" of the piece, their expressive microstructure would probably be just as different as it was before the tempi were equalized. In other words, tempo as such probably accounts only for a very small proportion of the variance among different performers' interpretations. Or, to put it yet another way, interpretation usually influences tempo, but tempo does not necessarily influence interpretation.

It will of course be desirable to employ a larger number of musicians in future studies, as well as a better instrument. The present two pianists' performances, however, seemed adequate enough for the purpose of the study. The fact that one pianist (BHR) was not professionally trained is not perceived to be a problem in view of the fact that his data were generally more consistent and less variable than those of the professional pianist, LPH. LPH's greater variability may be attributed to three sources: First, she had less time to adapt to the instrument; second, she had not played "Träumerei" recently; and third, as a professional artist she was not obliged to obey the constraints of a quasi-experimental situation. BHR, on the other hand, was more familiar with the Roland keyboard, had practiced Schumann's "Kinderszenen" less than a year ago, and, as an experimental psychologist, was used to exhibiting consistent behavior in the laboratory.[17]

In summary, the present findings suggest that for certain types of music it is possible to change the tempo within reasonable limits, either naturally in performance or artificially in a computer, without changing the quality and expression of a performance significantly. This suggests that the very complex motor plan that underlies artistic music performance contains something like a variable rate parameter that permits it to be executed at a variety of tempi while serving the same underlying cognitive structure. Various small-scale performance details, however, may be locally controlled and unaffected by the rate parameter. Thus local constancy appears to be nested within global flexibility in music performance.

REFERENCES

Clarke, E. F. (1982). Timing in the performance of Erik Satie's 'Vexations'. *Acta Psychologica*, 50, 1-19.

Clarke, E. F. (1985). Structure and expression in rhythmic performance. In P. Howell, I. Cross, & R. West (Eds.), *Musical structure and cognition* (pp. 209-236). London: Academic Press.

Clynes, M., & Walker, J. (1986). Music as time's measure. *Music Perception*, 4, 85-120.

Desain, P., & Honing, H. (1992a). Tempo curves considered harmful. In P. Desain & H. Honing (Eds.), *Music, mind, and machine* (pp. 25-40). Amsterdam: Thesis Publishers.

Desain, P., & Honing, H. (1992b). Towards a calculus for expressive timing in music. In P. Desain and H. Honing (Eds.), *Music, mind, and machine* (pp. 175-214). Amsterdam: Thesis Publishers.

Desain, P., & Honing, H. (in press). Does expressive timing in music performance indeed scale proportionally with tempo? *Psychological Research.*

Gentner, D. R. (1987). Timing of skilled motor performance: Tests of the proportional duration model. *Psychological Review, 94,* 255-276.

Guttmann, A. (1932). Das Tempo und seine Variationsbreite. *Archiv für die gesamte Psychologie, 85,* 331-350.

Handel, S. (1986). Tempo in rhythm: Comments on Sidnell. *Psychomusicology, 6,* 19-23.

Heuer, H. (1991). Invariant relative timing in motor-program theory. In J. Fagard & P. H. Wolff (Eds.), *The development of timing control and temporal organization in coordinated action* (pp. 37-68). Amsterdam: Elsevier.

MacKenzie, C. L., & Van Eerd, D. L. (1990). Rhythmic precision in the performance of piano scales: Motor psychophysics and motor programming. In M. Jeannerod (Ed.), *Attention and performance XIII* (pp. 375-408). Hillsdale, NJ: Erlbaum.

Michon, J. A. (1974). Programs and "programs" for sequential patterns in motor behavior. *Brain Research, 71,* 413-424.

Palmer, C. (1989). Mapping musical thought to musical performance. *Journal of Experimental Psychology: Human Perception and Performance, 15,* 331-346.

Rasch, R. A. (1988). Timing and synchronization in ensemble performance. In J. A. Sloboda (Ed.), *Generative processes in music* (pp. 70-90). Oxford, UK: Clarendon Press.

Repp, B. H. (1990). Patterns of expressive timing in performances of a Beethoven minuet by nineteen famous pianists. *Journal of the Acoustical Society of America, 88,* 622-641.

Repp, B. H. (1992). Diversity and commonality in music performance: An analysis of timing microstructure in Schumann's "Traumerei." *Journal of the Acoustical Society of America, 92,* 2546-2568.

Repp, B. H. (1993). Some empirical observations on sound level properties of recorded piano tones. *Journal of the Acoustical Society of America, 93,* 1136-1144.

Repp, B. H. (in press). On determining the global tempo of a temporally modulated music performance. *Psychology of music.*

Rorem, N. (1983). Composer and performance. [Originally published in 1967.] In N. Rorem, *Setting the tone* (pp. 324-333). New York: Coward-McCann, Inc.

Schmidt, R. A. (1975). A schema theory of discrete motor skill learning. *Psychological Review, 82,* 225-260.

Seashore, C. E. (1938). *Psychology of music.* New York: McGraw-Hill. (Reprinted by Dover Publications, 1967).

Shaffer, L. H. (1980). Analysing piano performance: A study of concert pianists. In G. E. Stelmach & J. Requin (Eds.), *Tutorials in motor behavior* (pp. 443-455). Amsterdam: North-Holland.

Shaffer, L. H. (1981). Performances of Chopin, Bach, and Bartok: Studies in motor programming. *Cognitive Psychology, 13,* 326-376.

Shaffer, L. H. (1984). Timing in solo and duet piano performances. *Quarterly Journal of Experimental Psychology, 36A,* 577-595.

Shaffer, L. H. (1992). How to interpret music. In M. R. Jones and S. Holleran (Eds.), *Cognitive bases of musical communication* (pp. 263-278). Washington, DC: American Psychological Association.

Shaffer, L. H., Clarke, E. F., & Todd, N. P. (1985). Metre and rhythm in piano playing. *Cognition, 20,* 61-77.

Todd, N. P. McA. (1992). The dynamics of dynamics: A model of musical expression. *Journal of the Acoustical Society of America, 91,* 3540-3550.

Vernon, L. N. (1937). Synchronization of chords in artistic piano music. In C. E. Seashore (Ed.), *Objective analysis of musical performance* (pp. 306-345). University of Iowa Studies in the Psychology of Music, Vol. 4. Iowa City: University of Iowa Press.

Viviani, P., & Laissard, G. (1991). Timing control in motor sequences. In J. Fagard & P. H. Wolff (Eds.), *The development of timing control and Temporal Organization in Coordinated Action* (pp. 1-36). Amsterdam: Elsevier.

FOOTNOTES

**Psychological Research,* in press.

[1]For temporal intervals to be proportional at different tempi, their absolute differences must be larger at slow than at fast tempi. Clarke mentions that an initial analysis yielded "identical" results for raw and log-transformed data, but it is not clear whether that was also true in the later analysis just referred to.

[2]The pianists' accuracy in implementing the desired tempi will not concern us here; this issue is the subject of a separate paper (Repp, in press). Naturally, the three performances within each tempo category were not exactly equal in tempo.

[3]These values represent the grand mean effects in one-way ANOVAs on the average scores minus 50%, equivalent to one-sample *t*-tests. The second test was conducted on the average scores of pairs of pairs containing the same type of comparison in opposite order (e.g., LPH's first medium-tempo performance followed by the speeded-up version of her first slow performance, and the speeded-up version of her second slow performance followed by her second medium-tempo performance).

[4]All correlations reported include multiple data points for IOIs longer than one eight-note (cf. Figure 2), which is equivalent to weighting longer IOIs in proportion to their duration. This seems a defensible procedure in comparing profiles, and omission of these extra points had a negligible effect on the magnitude of the correlations. However, in the ANOVAs reported further below, only a single data point was used for each long IOI, so as not to inflate the degrees of freedom for significance tests.

[5]Two differences between LPH and BHR are worth noting here: LPH prolonged the interval preceding the *fermata* chord much more than did BHR, and she did not implement the gestural boundary (i.e., the "comma" in the score) in bar 24, executing a continuous *ritardando* instead. Contrary to the score, LPH generally arpeggiated the chord under the *fermata;* hence the long IOI, which was measured to the last (highest) note of the chord.

[6]One possible reason for why LPH showed no significant interactions is that her tempo variation within tempo categories was larger than that of BHR (see Repp, in press), so that the separation of within- and between-tempo variation was less clear in her data.

[7]In bar 8, BHR showed a triple interaction of rendition (first vs. second playing), IOI, and tempo [$F(4,12) = 7.73$, $p < .003$], due to more even timing of the grace note in the first than in the second rendition, at the slow tempo only.

[8]BHR showed a marginally significant (but not easily characterizable) triple interaction with rendition in bar 6 [$F(4,12) = 3.42$, $p < .05$].

[9]In terms of the IOI ratios plotted in Repp's (1992) Figure 11, the respective coordinate values are approximately 0.5 and 0.6 for LPH, and 0.25 and 0.45 for BHR.

BEST COPY AVAILABLE

[10]The measurements were coarse because the 5-ms temporal resolution was poor relative to the small size of these intervals. However, the availability of repeated measurements within and across performances attenuated this problem considerably.

[11]It is common in piano performance for the pedal to be used more frequently than indicated in the score. It also should be duly noted that the pedal in this instance was a simple foot switch that did not permit the subtlety of pedalling possible on a real piano.

[12]LPH's pedalling times were much more variable than BHR's; this was also true within each tempo. The only substantial deviation for BHR occurred in bar 1, where pedal onset occurred at the beginning of the IOI in the absence of a preceding pedal offset. (In two instances at the medium tempo, however--not shown in the figure--there was a preceding pedal depression, and offset and onset times were similar to those in other bars.) Because of this deviation, BHR's bar 1 was omitted from further analyses of pedalling times. BHR never had a pedal offset near or within the IOI in bars 1R or 9, whereas LPH did some of the time; however, her following pedal onset times did not seem to depend on whether there was a pedal offset nearby.

[13]The intensity profiles of the expert performances studied by Repp (1992) are not available for comparison, as they are very difficult to determine accurately from acoustic recordings.

[14]This may be true only for the melody notes analyzed here, which indeed seemed to "stand out" more from among the four voices in LPH's performance.

[15]Surprisingly, however, there was a significant tempo main effect for BHR [$F(2,6) = 12.78$, $p < .007$], suggesting that the within-tempo changes in overall intensity evident in Figure 9 developed only later during his performances. For LPH, on the contrary, the between-tempo variation was conspicuously smaller than the within-tempo variation [$F(2,6) = 0.05$, $p > .95$] during the initial 8 bars.

[16]Their data have been augmented in the meantime and are presented in (Desain & Honing, in press).

[17]For those who are suspicious of authors serving as their own subjects, it might be added that BHR played with music, not hypotheses, in his mind.

Haskins Laboratories Status Report on Speech Research
1993, SR-113, 197-208

A Review of *Psycholinguistic Implications for Linguistic Relativity: A Case Study of Chinese* by Rumjahn Hoosain*

Yi Xu†

The content, and indeed the title, of this book brings back to us an old and well-known theory—the Sapir-Whorf hypothesis: that language influences thought and determines the world view of the speaker (Whorf, 1956). Rumjahn Hoosain proposes that "unique linguistic properties of the Chinese language, particularly its orthography, have implications for cognitive processing."

Although it does not seem to be his intention to revive the Sapir-Whorf hypothesis in its original form, Hoosain clearly displays his desire to salvage the theory through extension or revision. This is demonstrated by his remark in Chapter 1 that "the influence of orthography on cognition, particularly the influence of the manner in which the script represents sound and meaning, was not originally recognized in the work of Sapir and Whorf." Thus, Hoosain seems to be trying to keep the spirit of the Sapir-Whorf hypothesis alive while applying it more specifically to the effect of the orthography and particularly to the case of Chinese. This effect is, as he describes it, "more in terms of the facility with which language users process information, and the manner of information processing." Therefore, "such language effects matter more in data-driven or bottom-up processes rather than in conceptually driven or top-down processes."

Though not a large volume (198 pages), the book covers an impressively large number of psycholinguistic studies related to the Chinese language[1] and orthography, especially those presented at five conferences held in Taiwan and Hong Kong in the last dozen years. The reviews of those studies are spread over six chapters in the book, including an introduction in Chapter 1 and a conclusion in Chapter 6. The second chapter provides an introduction to general aspects of the Chinese language; and the third, fourth, and fifth chapters focus on, respectively, perceptual, memory-related, and neurolinguistic aspects.

Chapter 2 provides background information about the Chinese language, especially about its orthography. This chapter is necessary because there have been many myths and misconceptions about Chinese and its orthography that need to be clarified. Hoosain's description of Chinese, in general, is reasonable. However, there are certain remarks that are misleading and are still characteristic of common misunderstandings.

In 2.3, for example, Hoosain states that "a small percentage of the (Chinese) characters convey meaning by pictographic representation, either iconic or abstract," though he qualifies this by noting that "whereas these pictographic characters have an etymology related to pictures, this relation is unlikely to have psychological reality in present day usage." The critical issue here is the use of the phrase "convey meaning by pictographic representation." What a character in Chinese represents is a linguistic unit, in the case of the character for *niao3* [2] ("bird," used by Hoosain as an illustration of pictographic characters), a monosyllabic word. The word is a unit in the language—an entry in the lexicon. A lexical entry carries meaning, whereas a character is nothing but a visual symbol created and used to refer to a lexical item. The character by itself carries no meaning whatsoever. It conveys meaning only through its reference to the linguistic unit it represents. Thus, the character for *niao3* conveys the concept of "bird," not because it has the shape of a bird (otherwise, any

I am grateful to Ignatius G. Mattingly, Alvin M. Liberman, and Michael T. Turvey for their comments on an earlier version of this review.

picture of a bird would do), or because it is associated with the idea "bird," but because that shape is agreed by convention of the Chinese literate community to refer to a spoken word with the pronunciation /niao3/ (in present day Mandarin), which in turn refers to the concept of a feathered animal usually with flying ability.[3] On the other hand, real pictures of birds, whether vivid or sketchy, although they may well convey the meaning of bird directly, are not part of the orthography, and thus are completely different entities from the bird-shaped character for the word *niao3* . The character for *niao3* is (or has been) "pictographic" only in the sense that it was created by depicting the shape of a bird, and the pictographic structure of the character may have served as a mnemonic for the word. So, what can never be overemphasized is the fact that the bird-shaped character represents a LINGUISTIC unit—the spoken word for "bird" in Chinese. Therefore, a more accurate definition for so-called pictographic characters would be: They are characters created to represent spoken names of objects by depicting the visual pattern of the object the spoken names refer to. The shapes of the characters, when they are still reminiscent of the object the corresponding spoken words refer to, can be used as mnemonics for the spoken words.

In 2.5, while rejecting as a misconception the claim that characters in Chinese are logograms in the sense that each of them represents a whole word, Hoosain insists that "the meaning of a character is represented by the character directly, not mediated by sound symbols, and a syllable is indicated as a whole rather than spelled out, as in alphabetic languages." There are two confusions in this statement. First, as discussed above, Hoosain takes orthographic symbols as the meaning-carrying units by themselves. Second, by stating that "a syllable is indicated as a whole rather than spelled out," Hoosain sounds as if phonemes were the only possible meaningless linguistic units that can be represented by orthographic symbols. Apparently, Hoosain does not realize (or often forgets) that syllables are also meaningless units that can be used to construct (i.e., to spell) meaningful morphemes and words, and it is the syllable sign that is the basic graphemic unit in the Chinese orthography (DeFrancis, 1984, 1989; Mattingly, 1992). Instead, Hoosain takes the characters as the smallest graphemic unit in the Chinese writing system. As pointed out by DeFrancis (1989), the basic graphemes in Chinese are characters that singly represent whole syllables. They may themselves constitute frames, i.e., graphic units that are separated by white spaces in writing, or combine with other nonphonetic elements to form more complex characters constituting frames. Those graphemes, i.e., characters that provide phonological values when combining with other graphic elements to form other characters, are usually called phonetics, or phonetic radicals.

In his discussion in 2.4, Hoosain reports that there are about 800 phonetics in Chinese, and about 90% of Chinese characters are phonetic compounds. But in the discussion that follows, he belittles the function of the phonetics in Chinese orthography by emphasizing the irregularities and unreliabilities of their phonetic values, and by citing studies that show the semantic cueing function of the radical to be greater than the phonemic cueing function of the phonetic. It is not made clear by Hoosain what exactly is meant by "cueing function" in the unpublished study he cites. The other two studies (Yau, 1982; Peng, 1982) Hoosain cites show that semantic radicals tend to be located at the left or the top of the characters. "Because," Hoosain argues, "character strokes tend to be written left to right and top down, the beginning strokes therefore are more discriminative." Being more discriminative, however, is different from being more crucial. As a matter of fact, the common use of the word 'determinative' to refer to the semantic radicals in Chinese characters implies that their function is to help distinguish between characters that are otherwise homophones because they share the same phonetic radical. The premise here is that the phonological values of the phonetic radicals are already being used to the fullest possible extent. Of course, there do exist many irregularities and unpredictabilities in the phonological values of the phonetic radicals in Chinese. What is wrong in Hoosain's account of the Chinese orthography is that he is using those irregularities to deny the basic organization principles in the construction of the characters. This is like denying the phonemic principle in the construction of English words merely because there are tremendous irregularities in the letter-phoneme correspondence.

Psychological studies of languages and reading can not and should not be totally independent of the knowledge gained over the years about the basic principles of language and speech and of writing systems. Hoosain's failure to fully appreciate those principles is reflected in his view that Chinese characters represent both meaning

and sound directly, which is best illustrated in the diagram on page 12 of his book.

Chinese: sign (script) ↗ sound / ↘ meaning

English: signs (script) → sound → meaning

Notice here his ambiguous use of the words 'sign' and 'sound.' Apparently, by 'sign,' he means letters for English but characters for Chinese. By 'sound,' he seems to mean phonemes for English but syllables for Chinese. As discussed above, while letters are the smallest graphemic units in English, characters are not the smallest graphemic units in Chinese. The smallest graphemic units in Chinese are the phonetic radicals in characters.[4] So, it is the phonetic radicals that map onto syllables directly (although no more accurately than letters map onto phonemes in English), not the characters. Likewise, it is the morphemes that characters map onto, not sound.

As we will see, the confusion in the understanding of the linguistic and orthographic aspects of the Chinese language is bound to affect Hoosain's interpretation of the studies he reviews.

Chapter 3 reviews studies concerned with perceptual aspects of the Chinese language. The conclusion reached by Hoosain in this chapter is that "the distinctive configuration of the Chinese script as well as its script-sound and script-meaning relation can differentially affect perceptual processes, compared with similar processes in English." The effects include: "(a) lack of acuity difference for characters arranged in different orientations; (b) preference of direction of scan according to language experience; (c) the individuality of constituent characters of disyllabic words in some situations; (d) the more direct access to meaning of individual words, although access to sound could require a different effort due to the lack of grapheme-phoneme conversion rules; and (e) the variation in eye movements in reading connected with different manners of reading." Due to lack of space, and because they are not essential concerns in the study of reading, the first two effects and the last effect will not be discussed in this review. The only thing that needs to be pointed out is that letters in English and characters in Chinese are at different levels of linguistic representation in the two orthographies, and, therefore, cannot be equated. In other words, differences found between the two may not always unambiguously point to critical differences between the reading mechanisms of the two languages.

When discussing the size of perceptual units in reading, it is essential to make sure the units being discussed are unambiguous. That is to say, when characters are referred to, we should not sometimes mean words and sometimes morphemes, and we should not equate them with other units, say, letters. This is not always easy, because in Chinese morphemes and words often coincide. This difficulty, however, makes it all the more important for a reviewer of reading studies to watch out for dubious hidden assumptions ignoring this distinction, and for flawed designs as well as improper interpretations based on those assumptions. This is, however, exactly what Hoosain often fails to do as a reviewer.

For example, in his discussion of the study by C. M. Cheng (1981) on word superiority, Hoosain concludes that characters in disyllabic Chinese words have more perceptual salience than letters in English, and he calls this phenomenon a "contrast between the two languages." As discussed above, characters in Chinese correspond to morphemes, and morphemes are the smallest meaning carrying units in languages. Letters in English, however, correspond to phonemes, i.e., the smallest distinctive segments in speech, though the correspondence is often inconsistent. Phonemes do not carry meaning by themselves. They have to combine with other phonemes to form meaning-carrying morphemes. So, perceptual differences found between characters in Chinese and letters in English should be regarded as a reflection of the differences between the two types of linguistic units tested, rather than as an indication of any true difference between the psycholinguistic processing of the two languages.

Liu (1988) found that a Chinese character standing alone was named as fast as the same character occurring as a constituent in a two-character word. In contrast, a simple word in English, such as *air*, is named slower in isolation than as a constituent of a compound word, such as *airways*. Hoosain interprets these findings as an indication that "constituent characters of Chinese words are handled somewhat independently, rather than as integral parts of words." However, there are several problems with Liu's study. First, Liu did not mentioned whether the constituent characters of his Chinese stimuli could also function as words by themselves. (No sample list of the Chinese words used was provided.) Second,

neither word frequency in English and Chinese nor character frequency in Chinese was controlled in the study, which makes any differences found between naming latencies of compound words and their constituents suspect. Third, native speakers of Chinese who had also learned to speak English were used as subjects for both Chinese and English. This use of non-native speakers of English for reading English in comparison with native speakers of Chinese reading Chinese is problematic. Fourth, as Hoosain points out himself, the English compounds like *baseball* were all presented with a hyphen in the middle, no matter whether they are conventionally printed that way or not. There is no telling what effect this arbitrary manipulation may have had. So, the results of this study as evidence of "another difference between Chinese and English" are, to say the least, inadequate.

In summarizing the section on perceptual units, Hoosain concludes that "particularly in cases where component characters are simple and highly familiar, or when individual characters have to be pronounced, there is some degree of individual salience for these characters that is not likely to be found in the case of bound morphemes in English. This individual salience has to do with the character as sensory unit." While the statement about the salience of individual characters may be true to a certain extent, the comparison between characters in Chinese and bound morphemes in English is inappropriate. In Chinese, most of the high frequency characters can function as monosyllabic words by themselves. Of the 1185 characters with frequency of occurrence of 100 or higher per million characters, only 44, i.e., 3.7%, can not function as monosyllabic words (Wang et al., 1986). The only comparable situation in English is in cases where words are also used as constituents in compound words. But the only study Hoosain cites that made this comparison is flawed. Without better evidence, therefore, the best conclusion that can be drawn about this matter is that further studies are needed.

In the discussion in 3.3 on getting at the sound of words in Chinese, Hoosain insists that "the only way to get at the sound of Chinese characters is through lexical information, whether directly or indirectly." His major argument is that in Chinese characters, "the phonetics are characters on their own, and their own pronunciations have to be learned on a character-as-a-whole-to-sound-as-a-whole basis and not through grapheme-phoneme conversion rules." Here Hoosain seems to confuse the difference in level of representation of speech sound with the presence or absence of grapheme-speech-sound correspondence. The lack of grapheme-phoneme correspondence in Chinese does not mean the lack of grapheme-speech-sound correspondence altogether. Graphemes in Chinese correspond to speech sound at the level of the syllable (DeFrancis, 1989; Mattingly, 1992). This correspondence is not one to one, but neither is the correspondence between letters and phonemes one to one in English. This confusion between levels of representation and existence of grapheme-speech-sound correspondence determines Hoosain's bias when reviewing studies on phonological access in reading Chinese.

In examining studies on naming latency for characters, Hoosain focuses on Seidenberg (1985). In that study, it was found that naming times were faster for phonetic compounds than for nonphonetic compounds, but only for low frequency characters and not for high frequency characters. Seidenberg also did a parallel study with English, in which he found that words with exceptional pronunciations were named slower than those with regular pronunciations. But again, this was found to be true only for low frequency words. Seidenberg took these results to mean that, for both Chinese and English, access to the pronunciation is through the lexicon for high frequency words and through grapheme-to-sound conversion for low frequency words.

Judging from the example given in Seidenberg's paper, there are problems in the material he used in his Chinese experiment. Both of the so-called nonphonetic compound characters can be used as phonetic radicals in many other characters. So, they are in principle also phonetic compounds, except that their semantic radicals are nil. Besides, Seidenberg did not control the frequency or the phonological consistency of the phonetic radicals used in his study. Those problems makes Seidenberg's finding questionable.

Hoosain also notices some of the above mentioned problems in Seidenberg's study, but he questions the results from a different direction. He argues that access to whatever phonological information is provided by the phonetic radicals in Chinese cannot be compared to grapheme-phoneme conversion in alphabetic languages. And he raises an intriguing question: "How can it be argued that access to the sound of such nonphonetic compounds (when they stand alone) is through lexical information, but access to the sound of phonetic compounds (which are actually made up of these same nonphonetic compounds

now acting as phonetics) is through graphemic-to-phoneme conversion? After all, how is the sound of these phonetics arrived at in the first place?"

While this question is well posed, Hoosain's answer is problematic. He insists that access to the pronunciations of all the characters in Chinese is through the lexicon, either directly or indirectly. This would imply that the naming latency should be the same whether the phonetic radicals have consistent or inconsistent phonological values in the characters that they are part of. However, this prediction does not agree with the findings by Fang, Horng, and Tzeng (1986), which is also mentioned by Hoosain. They demonstrated that high consistency in pronunciation shared among characters with the same phonetic radical facilitated naming latency for simple characters as well as for compound characters and pseudocharacters. This consistency effect was shown to be comparable to a similar effect found in naming latency for English words (Glushko, 1979).

Another phenomenon taken seriously by Hoosain is the minor finding by Seidenberg (1985) that the overall naming latency was much slower for his Chinese subjects than for his English subjects. Hoosain takes this as an indication that phonological access in Chinese is slower than in English. It is not noticed by Hoosain, however, that the English subjects Seidenberg used were college students, whereas the Chinese subjects he used were described only as native Cantonese. This makes one wonder what their educational backgrounds were and even how much reading experience they had had. Undoubtedly, these factors would have effectively influenced their naming latency.

To summarize the discussion of getting at the sound of words, there is so far no solid evidence for Hoosain's view that phonological access for characters in Chinese is solely through the lexicon. The studies by Seidenberg (1985) and by Fang et al. (1986) suggest that more evidence for the use of phonetic components in characters in getting at the sound of characters is very likely to be found.

In 3.4, Getting at the Meaning of Words, Hoosain claims to have found evidence that getting at the meaning of Chinese characters is more direct than is the case with printed words in alphabetic writing systems. At the outset, Hoosain rightly warns the reader to bear in mind that "spoken language is prior to written language... Therefore sound-meaning relation is prior to script-meaning relation, giving rise to the question of the need for phonological recoding before access to meaning." By "prior," however, Hoosain means the order in historical development, both for human languages in general and for individual speakers of a language in particular, leaving plenty of room for the actual reading process to either involve or not involve so-called phonological mediation. This leads to the next question he asks, namely, "whether phonological recoding is involved to a different extent depending on the nature of the orthography." So, it follows, according to Hoosain, that "if the script primarily represents sounds of the language, script-meaning relation may not be so direct, and getting at the meaning more likely may be mediated by phonological recoding."

With these questions in perspective, Hoosain presents several studies he thinks show a more direct connection between script and meaning in Chinese than in alphabetic orthographies. The first series of studies involve the famous Stroop phenomenon. Stroop (1935) found that when the name of a color was written in ink of another color, it took subjects longer to name the color of the ink than it took them to name the colors of ink patches. This phenomenon was taken to mean that somehow it is unavoidable to process the meaning of the printed words even when the task is to attend to the color of the ink. This explanation, however, is questionable, since the phenomenon can also be understood as an indication that it is somehow unavoidable to process the *pronunciation* of the printed words, thus slowing down the naming process when there is disagreement in pronunciations between the printed color name and the color of the ink. With this alternative explanation, the even greater Stroop effect for Chinese color names found by Biederman and Tsao (1979) could, contrary to Hoosain's interpretation, be interpreted as an indication that it is even more unavoidable to process the pronunciation of Chinese characters. Without further evidence, though, the implication of the greater Stroop effect for Chinese should at least be considered undetermined.

In trying to find more evidence for direct character-meaning connection in Chinese, Hoosain cites the study by Tzeng and Wang (1983), but misunderstands the procedure used in that study.[5] In the study, subjects were asked to choose from two numbers written in characters the one that is numerically greater than the other. It was found that it took the subjects longer to make the decision when the characters for the larger number had a smaller physical size on the

screen. Hoosain, however, mistakenly describes the task as that of deciding on the physical size of the written numbers. He thus interprets the results as indicating that it is unavoidable to process the meaning of the Chinese numbers when the task is only about the physical size of the printed numbers. This interpretation is inappropriate because the actual task in the experiment required subjects to focus on the meaning of the numbers to begin with.

As for the finding that Chinese numbers are translated faster into English than the other way around by native Cantonese speakers (Hoosain, 1986), which Hoosain uses as another piece of evidence for a direct connection between characters and meaning, I find it inconclusive because no comparable results were obtained from native speakers of English. Similarly, the finding that cross-language priming effects for native speakers of Chinese are greater from Chinese to English than in the other direction (Keatley, 1988) is not conclusive evidence for a direct connection between meaning and characters in Chinese.

Another study cited by Hoosain as showing evidence for a direct character-meaning connection was conducted by Treiman, Baron, and Luk (1981). In that study, native speakers of English and Cantonese were required to make truth/falsity judgments of sentences in their own languages. A sentence could be a "homophone" sentence, such that it sounded true when read out loud (e.g., *A pair is a fruit*). It was found that it took longer to reject these homophone sentences than truly false sentences like "A pier is a fruit." In contrast, this kind of delay due to homophony was relatively small for Chinese, and this difference was taken as evidence that phonological recoding was less involved in Chinese sentence processing. However, there is a problem in that study that Hoosain does not notice. In the sample Chinese sentence presented in the paper, *Mei2 shi4 yi1 zhong3 zhi2 wu4* [6] ("*Mei2* is a plant"), the character *mei2*, meaning "coal," is said to be homophonous with the name of a plant. But the character for *mei2* as name of that plant is not a meaningful word in itself in the spoken language (neither in Cantonese nor in Mandarin). In the spoken language, it always combines with some other syllables to form words. The syllable *mei2* by itself corresponds only to one word, namely, "coal." So, the smaller impairment for the Cantonese speakers in that experiment may well be a consequence of using improper test material rather than an indication of less phonological recoding in Chinese sentence processing.

The only study that offers a somewhat stronger case for possible fast access to the meanings of Chinese characters is the one by Hoosain and Osgood (1983) on affective meaning response times. They asked subjects to report the affective meaning of a word by saying either "positive" or "negative" upon seeing the word displayed on a screen. Native speakers of Cantonese could make this judgment faster than native speakers of English. Hoosain and Osgood interpreted this result as evidence that the processing of at least some aspects of meaning of Chinese words was faster than that of English words, and suggested that this was because phonology was bypassed when accessing the meaning of printed Chinese words. Fascinating as the finding may be, it is somewhat in conflict with a recent study by Perfetti and Zhang (1991, also see Perfetti, Zhang, and Berent, in press), in which it was found that phonemic information was immediately available as part of character identification in Chinese. Furthermore, these recent experimental results suggest that whatever the semantic value of a character, it is activated no earlier than its phonological value. If the findings in both studies (Hoosain and Osgood, 1983; and Perfetti and Zhang, 1991) are true, then the faster affective meaning response time for Chinese words should only mean faster decision after the activation of the phonology as well as the semantics of those words, rather than faster activation of the semantics, assuming, of course, that the phonological activation of the Chinese words is no faster than that of the English words.

In conclusion, I do not find solid evidence for a more direct connection between meaning and characters in Chinese than between meaning and printed words in alphabetic orthographies. Nor do I find solid evidence for lack of phonological mediation in accessing meaning from characters in Chinese. Rather, the studies by Seidenberg (1983) and Fang et al. (1986), which are also cited by Hoosain, show evidence of similarity between the processing of Chinese orthography and the processing of alphabetic orthographies.

Chapter 4 of Hoosain's book focuses on memory-related aspects of the Chinese language. In general, Hoosain shows in this chapter that phonological recoding is used in memorizing printed Chinese, but, at the same time, he cites some studies that he thinks show a greater role of visual coding for Chinese than for other languages.

In 4.1.1, Hoosain cites quite a few studies that found a greater digit span for Chinese than for

some other languages. It is shown that this difference can be well accounted for by the working memory model proposed by Baddeley and Hitch (1974) and Baddeley (1983, 1986). This model assumes an "articulatory loop mechanism" for short-term memory which has a span of approximately two seconds worth of speech sound. Hoosain shows that number names in Chinese have shorter duration in pronunciation than in some other languages. Because of this, Hoosain argues, more numbers can be kept in the articulatory loop for the Chinese speakers.

The discussion in this section is interesting, except that there is no emphasis on the fact that the phenomenon is purely linguistic, and has almost nothing to do with orthography. Also, Hoosain does not emphasize that the finding is interesting because it confirms one of the hypothetical principles under which all human languages function (but see Ren and Mattingly, 1990; Mattingly, 1991, for criticism of the working memory hypothesis). Instead, Hoosain seems to be thrilled by the finding itself, that digit span in Chinese is larger than in some other languages, and uses this as an indication that differences among languages do affect the way information is processed.

In his discussion of the visuo-spatial scratch-pad and perceptual abilities, Hoosain cites a study by Woo and Hoosain (1984). They conducted a short-term memory test, in which subjects had to identify among a list of seven characters those that had appeared among the six characters shown to them a few seconds before. It was found that "weak readers made much more visual distracter errors (when target characters were mixed with similar looking distracters) but not more phonological distracter errors (when targets were mixed with similar sounding distracters)." Hoosain contrasts this finding with that of Liberman, Mann, Shankweiler, and Werfelman (1982), who found that, for American subjects, good beginning readers of English did not do better than weak beginning readers in memory for abstract visual designs or faces. However, they did perform better in memorizing nonsense syllables. In comparing this result with the results of Woo and Hoosain (1984), Hoosain comes to the conclusion that poor readers in Chinese have more visual problems than poor readers in English. Apparently, Hoosain misinterprets the nature and the implication of the results obtained by Woo and Hoosain. The finding that poor readers made more visual distracter errors indicates that they were relying more heavily than good readers on the visual characteristics of the characters when trying to encode them for immediate recall. As a result, poor readers were penalized more by the graphic similarities brought about by the visual distracters. The finding does not, as Hoosain interprets it, simply indicate that poor readers in Chinese have more visual problems. In a comparable study by Ren and Mattingly (1990), it was found that good readers in Chinese were relatively more affected by phonologically similar series than poor readers when performing immediate recall of series of Chinese characters. These results are comparable to those found for English readers (Shankweiler, Liberman, Mark, Fowler, and Fischer, 1979). In neither of those two studies, the results were interpreted by the authors as evidence that good readers had more phonological problems than poor readers. Rather, they were taken to suggest that good readers rely more heavily than poorer readers on phonological encoding, hence were penalized more by the phonological similarity introduced in the experimental stimuli.[7] So, the results of Woo and Hoosain (1984) and Ren and Mattingly (1990) both show that one of the crucial differences between good and poor readers in Chinese is in the manner in which they recode reading material; Good readers use more phonological encoding while poor readers rely more on visual encoding. This agrees, rather than contrasts, with findings about the differences between good and poor readers in English (Shankweiler et al.).

In a study by Chen and Juola (1982), native Chinese and English speaking subjects had to decide which of the items on separate test lists was graphemically, phonemically or semantically similar to the item just presented visually. Chinese subjects were found to be fastest in making graphic similarity decisions about characters while American subjects were found to be fastest in phonemic similarity judgments. Chen and Juola then concluded, as reported by Hoosain without reservation, that "Chinese words produce more distinctive visual information, but English words result in a more integrated code." Thus the two different "scripts activate different coding and memory mechanisms." Hoosain, however, does not notice at least two problems with the study. First, the study claimed to use words as stimuli both for English and Chinese. However, of the 144 characters used, 26 could function only as bound morphemes according to the frequency dictionary of Wang et al. (1986), and another 11 could not even be found in that dictionary. Using bound morphemes in Chinese necessarily involves more

homophones, making graphemic discriminations more indispensable. Second, the task the subjects had to accomplish requires the kind of phonological awareness that most Chinese readers do not need in ordinary reading. Yet, lacking conscious knowledge about the phonemic structure of characters does not necessarily mean the absence, or even less use, of phonological recoding in actual reading. So, here Hoosain is confusing phonological recoding in actual reading with conscious phonological manipulation in performing certain cognitive experimental tasks.

Another study Hoosain cites as evidence for existence of a visual coding strategy for Chinese materials is that of Mou and Anderson (1981). These authors found predominant use of phonological recoding in STM for Chinese. Yet, at the same time, they found evidence of visual recoding. In that study, the presentation of stimulus lists was simultaneous rather than serial. As pointed out by Yu, Jing, and Sima (1984) and Yu, Zhang, Jing, Peng, Zhang, and Simon (1985), measured STM capacity varies depending upon the visual presentation method, i.e., visual aspects of the stimuli have greater effect upon the results with simultaneous presentation than with serial presentation. Yu et al. (1985) also concluded, based on the results of 14 experiments they carried out in China and the USA, that "no effects were detected that are peculiar to ideographic or logographic languages in contrast to alphabetic languages."

In 4.4 Hoosain looks at memory in relation to reading problems in Chinese. He reports, in particular, an extensive study conducted by Stevenson, Stigler, Lucker, Lee, Hsu, and Kitamura (1982) that compared reading disability among Chinese, American, and Japanese children. The major finding was that the proportion of children with reading problems is very similar among the three different languages, thus disconfirming the previous popular belief that there are fewer reading problems among Chinese and Japanese children. The study did find, however, an interesting difference between Chinese and American as well as Japanese children. Of those Chinese children who failed the fifth grade test, most failed because of a poor comprehension score, whereas most of the fifth grade American or Japanese children who failed the test did so because of their poor vocabulary score. Hoosain suggests this is because pronunciation tests involve more straightforward answers than word meaning tests, whereas a more open-ended effort is required for questions about meaning. He further suggests in 4.5, that because of the nature of the Chinese orthography, Chinese children have to do more rote learning. And, because this kind of learning calls for a large measure of sustained discipline, authoritarianism is thus tied to this system of learning, which in turn restrains students from giving more open-ended answers. This kind of reasoning, although interesting, is only one of the possible ways of account for the data. An alternative explanation is that the poor comprehension score for Chinese children may be due to improper testing design. Teaching of reading and writing in Chinese is in general character-oriented. However, characters correspond to morphemes, and most of the morphemes need to combine with other morphemes to form a word. So, knowing the pronunciation of a character and its meaning in one or several of the words that contain it does not necessarily lead to knowing its meaning in other words it is part of. This discrepancy between the target of teaching and the units of comprehension makes it difficult for people to design tests for measuring progress in comprehension, especially when it is assumed that knowing the pronunciation of all the characters in a text and their meaning in isolation should be sufficient for comprehension of the meaning of the text.

To summarize the discussion of Chapter 4, some of the alleged differences reported by Hoosain between the memory aspect of Chinese and other languages are found to be inappropriate interpretations of the experimental findings, while others derive from studies with defective designs. In conclusion, so-called bottom-up differences, or differences between processing of Chinese and other languages, may well be just a reflection of differences among languages themselves, or merely of the difficulty in finding parallel ways of measuring performance in different languages.

In Chapter 5, Neurolinguistic Aspects of the Chinese Language, Hoosain presents convincing evidence that the neurolinguistic aspects of Chinese are not essentially different from other languages, as suggested by some of the early studies. At the same time, however, there is a persisting confusion in Hoosain's review between language and orthography. Though this should not be entirely blamed on him, since most of the studies he reviews did not make this distinction in the first place, as a reviewer he should have been more aware of this confusion.

In 5.1, Hoosain shows that despite a claim (e.g., Hatta, 1977) that characters are processed in the right hemisphere, there are plenty of other studies that demonstrated left hemisphere advantage for

printed Chinese words as well as English words. It is further shown that right hemisphere advantage for Chinese was only found with single characters, whereas two-character words always elicited left hemisphere advantage. Also, with single characters the findings are mixed: Both right-hemisphere advantage and left-hemisphere advantage are found, and which hemispheric advantage is obtained seems to depend heavily on the test conditions. Right-hemisphere advantage seems to be related to short exposure time, low illumination, and high structural complexity of the characters. This is further confirmed by the finding (Ho and Hoosain, 1989) that, with short exposure time, low luminance, and high stroke number, right-hemisphere advantage could be observed even for low frequency two character words.

In 5.3 and 5.4, Hoosain reviews twenty-one studies on Chinese aphasics. His general conclusion is that "there is little evidence to suggest any overwhelming neurolinguistic effects of Chinese language uniqueness." Although it was suggested by some earlier studies that Chinese language functions are more lateralized in the right hemisphere, results of more extensive investigations indicate no such tendency. In general, the proportion of right-handed patients with aphasia after right hemisphere damage does not exceed the overall proportion of right handed people who have their speech functions lateralized in the right hemisphere (about 4%). So, Hoosain concludes that "it is quite clear from these studies that specialization for language tends to be in the left hemisphere for the Chinese, as for speakers of alphabetic languages." (Notice, however, his use of the word 'language' when most of the studies were actually about reading.)

There is an indication from some studies that Chinese reading functions may be localized more posteriorly, involving the parietal and occipital lobes, and psycho-motor schemas may play a greater role in memory for Chinese words. However, as Hoosain points out, it is premature to conclude that there is more involvement of the occipito-parietal region for Chinese. As for reports that finger tracing has been used by some patients to help recognize Chinese characters, which Hoosain regards as essentially different from spelling out individual letters for helping read words in an alphabetic orthography, it seems that those patients' reading ability has been impaired so much that they are doing only what beginning learners of an orthography are doing. It is very likely that both finger tracing and oral spelling reflect only the way word composition is learned for different orthographies,[8] and this process could be quite peripheral to the essential mechanisms involved in fluent reading.

The finding that processing of tones in Chinese is located in the left hemisphere is interesting, although again this is an aspect of language and is thus not directly related to orthography. A study by Packard (1986) showed that patients with left hemisphere damage had defective production of lexical tones in Mandarin. On the other hand, Hughes, Chan, and Su (1983) found that patients with right hemisphere damage had defects in both production and comprehension of affective intonation, but at the same time had kept their ability to process lexical tones almost intact. The finding that processing of tones is located in the left hemisphere, together with the finding that processing sign language (e.g., Damasio, Bellugi, Damasio, Poizner, and Gilder, 1986) is also located in the left hemisphere, reveals the universality shared by all human languages not only in terms of their ability to convey information, but also in terms of the central neural processing mechanism they all utilize.

To summarize this discussion of Chapter 5, the evidence Hoosain provides for similarity in lateralization of script as well as general linguistic processing in the brain between Chinese and other languages is convincing. The tendency for right hemisphere advantage in processing isolated characters shown in some of the studies reviewed by Hoosain, however, does need further exploration.

In final conclusion, Hoosain's major attempt in this book is to introduce a new version of the linguistic relativity hypothesis. The original version is the famous Sapir-Whorf hypothesis that language influences thought. The new version is concerned less with structure of language or world view than with the manner in which linguistic information is processed, and in particular, with the script-sound-meaning aspects of the Chinese orthography and their effects on information processing. Hoosain's final conclusion in his book is that "there are definite correlations between language characteristics and cognitive performance." However, I find that this conclusion is based mainly on misconceptions about the Chinese language and its orthography, on results obtained under defective experimental designs, and on misinterpretations of the results due to those misconceptions. Besides, the difficulty in separating language differences and orthographic differences from processing differences also affects Hoosain's

judgments. More specifically, the difficulty is in finding truly consistent measurements in comparing different languages and orthographies. Unless this difficulty is overcome, there is always the danger of mistaking the differences due to inconsistent measurements for true contrasts in the processing of different languages and orthographies.

So, as a final comment on the title of Hoosain's book, to parallel Hoosain's borrowing from physics the notion of 'relativity' to underscore his vision of linguistic processing of different languages, I would like to borrow, also from physics, the notion of 'uncertainty' or 'indeterminacy' to emphasize the caution we should exercise when comparing different languages and orthographies. It is often difficult to determine at the same time both the difference between two languages or writing systems and the differences in their processing. Comparing the processing of Chinese characters directly with the processing of letters, words, or bound or unbound morphemes in an alphabetic writing system is often risky; and, attributing any differences found in this kind of comparison to cognitive or behavioral differences will often prove to be a pitfall.

REFERENCES

Baddeley, A. D. (1983). Working memory. Philosophical Transactions of the Royal Society, London, B302, 311-324.

Baddeley, A. D. (1986). *Working memory*. Oxford: Clarendon.

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. A. Bower (Ed.), *The Psychology of Learning and motivation: Advances in research and theory*, Vol. 8 (pp. 47-90). New York: Academic Press.

Biederman, I., & Tsao, I. C. (1979). On processing Chinese ideographs and English words: Some implications from Stroop-test results. *Cognitive Psychology, 11*, 125-132.

Chen, H. C., & Juola, J. F. (1982). Dimensions of lexical coding in Chinese and English. *Memory & Cognition, 10*, 216-224.

Cheng, C. M. (1981). Perception of Chinese characters. *Acta Psychologica Taiwanica, 23*, 281-358.

Damasio, A., Bellugi, U., Damasio, H., Poizner, H., & Van Gilder, J. (1986). Sign language aphasia during left-hemisphere amytal injection. *Nature, 322*, 363-365.

DeFrancis, J. F. (1984). *The Chinese Language: Facts and Fantasy*. Honolulu: University of Hawaii Press.

DeFrancis, J. F. (1989). Visible speech: The diverse oneness of writing systems. Honolulu: University of Hawaii Press.

Fang, S. P., Horng, R. Y., & Tzeng, O. J. L. (1986). Consistency effects in the Chinese character and pseudo-character naming tasks. In H. S. R. Kao & R. Hoosain (Eds.), Linguistics, psychology, and the Chinese Language (pp. 11-21). Hong Kong: University of Hong Kong Centre of Asian Studies.

Glushko, R. J. (1979). The organization and activation of orthographic knowledge in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance, 5*, 674-691.

Hatta, T. (1977). Recognition of Japanese kanji in the left and right visual fields. *Neuropsychologia, 15*, 685-688.

Ho, S. K., & Hoosain, R. (1989). Right hemisphere advantage in lexical decision with two-character Chinese words. *Brain and Language, 27*, 606-615.

Hoosain, R. (1986). Psychological and orthographic variables for translation asymmetry. In H. S. R. Kand & R. Hoosain (Eds.), *Linguistics, psychology, and the Chinese Language* (pp. 203-216). Hong Kong: University of Hong Kong Centre of Asian Studies.

Hoosain, R., & Osgood, C. E. (1983). Information processing times for English and Chinese words. *Perception & Psychophysics, 34*, 573-577.

Hughes, C. P., Chan, J. L., & Su, M. S. (1983). Aprosodia in Chinese patients with right cerebral hemisphere lesions. *Archives of Neurology, 40*, 732-736.

Keatley, C. W. (1988). *Facilitation effects in the primed lexical decision task within and across languages*. Unpublished doctoral dissertation. University of Hong Kong.

Liberman, I. Y., Mann, V. A., Shankweiler, D., & Werfelman, M. (1982). Children's memory for recurring linguistic and non-linguistic material in relation to reading ability. *Cortex, 18*, 367-375.

Liu, I. M. (1988). Context effects on word/character naming: Alphabetic versus logographic languages. In I. M. Liu, H. C. Chen, & M. J. Chen (Eds.), *Cognitive aspects of the Chinese Language* (pp. 81-92). Hong Kong: Asian Research Service.

Mattingly, I. G. (1991). Modularity, working memory, and reading disability. In S. A. Brady & D. P. Shankweiler (Eds.), *Phonological processes in literacy: A tribute to Isabelle Y. Liberman* (pp. 163-171). Hillsdale, NJ: Lawrence Erlbaum Associates.

Mattingly, I. G. (1992). Linguistic awareness and orthographic form. In L. Katz & R. Frost (Eds.), *Orthography, phonology, morphology, and meaning* (pp. 11-26). Amsterdam: Elsevier Science Publishers.

Mou, L. C., & Anderson, N. S. (1981). Graphemic and phonemic codings of Chinese characters in short-term retention. *Bulletin of the Psychonomic Society, 17*, 255-258.

Packard, J. L. (1986). Tone production deficits in nonfluent aphasic Chinese speech. *Brain and Language, 29*, 212-223.

Peng, R. X. (1982). A preliminary report on statistical analysis of the structure of Chinese characters. *Acta Psychologica Sinica, 14*, 385-390.

Perfetti, C. A., & Zhang, S. (1991). Phonological processes in reading Chinese words. *Journal of Experimental Psychology: Learning, Memory and Cognition, 17*, 633-643.

Perfetti, C. A., Zhang, S., & Berent, I. (1992). Reading in English and Chinese: evidence for a "universal" phonological principle. In L. Katz & R. Frost (Eds.), *Orthography, phonology, morphology, and meaning*. Amsterdam: Elsevier Science Publishers.

Ren, N., & Mattingly, I. G. (1990). Short-term serial recall performance by good and poor readers of Chinese. *Haskins Laboratories Status Report on Speech Research. SR-103/104*, 153-164.

Seidenberg, M. S. (1985). The time course of phonological code activation in two writing systems. *Cognition, 19*, 1-30.

Shankweiler, D., Liberman, I. Y., Mark, L. S., Fowler, C. A., & Fischer, F. W. (1979). The speech code and learning to read. *Journal of Experimental Psychology: Human Learning and Memory, 5*, 531-545.

Stevenson, H. W., Stigler, G. W., Luker, G. W., Lee, S. Y., Hsu, C. C., & Kitamura, S. (1982). Reading disabilities: The case of Chinese, Japanese, and English. *Child Development, 53*, 1164-1181

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*, 643-662.

Treiman, R. A., Baron, J., & Luk, K. (1981). Speech recoding in silent reading: A comparison of Chinese and English. *Journal of Chinese Linguistics, 9*, 116-125.

BEST COPY AVAILABLE

Tzeng, O. J. L., Hung, D. L. , & Wang, W. S.-Y. (1977). Speech recoding in reading Chinese characters. *Journal of Experimental Psychology: Human Learning and Memory., 3,* 621-630.

Tzeng, O. J. L., & Wang, W. S.-Y. (1983). The first two R's. American Scientist. 71, 238-243.

Wang, H., Chang, B. R., Li, Y. S., Lin, L. H., Liu, J., Sun, Y. L., Wang, Z. W., Yu, Y. X., Zhang, J. W., & Li, D. P. (1986). *A Frequency Dictionary of Current Chinese.* Beijing: Beijing Language Institute Press.

Whorf, B. L. (1956). Language, thought, and reality. Selected writings of Benjamin Lee Whorf. Cambridge, MA: MIT Press.

Woo, E. Y. C., & Hoosain, R. (1984). Visual and auditory functions of Chinese dyslexics. *Psychologia, 27,* 164-170.

Yau, S. C. (1982). Linguistic analysis of archaic Chinese ideograms. *Paper presented at the XV International Conference on Sino-Tibetan Languages and Linguistics.* Beijing.

Yu, B., Jing, Q., & Sima, H. (1984). STM capacity for Chinese words and idioms: Chunking and acoustical loop hypotheses. *Memory & Cognition, 13,* 193-201.

Yu, B., Zhang, W., Jing, Q., Peng, R., Zhang, G., & Simon, H. A. (1985). STM capacity for Chinese and English materials. *Memory & Cognition. 13,* 202-207.

FOOTNOTES

Language and Speech, 35, 325-340 (1992).

†Also University of Connecticut, Storrs.

[1]Linguistically, Chinese is actually a large language family consisting of many mutually unintelligible languages. However, because the same writing system is used by the whole Chinese speaking community in China, and because of the close cultural ties among the Chinese people as well as the centralized political traditions, the Chinese languages are often considered dialects of a single language.

[2]The number 3 at the end of the spelling indicates the tone of the syllable.

[3]Incidentally, but perhaps more importantly, there are many different shades of meanings attributed to this word, as is the case with almost all the words in any language. This quality is apparently lacking in any true picture of a bird.

[4]Chinese character is typically composed of a phonetic radical and a semantic radical. Sometimes there is more than one semantic radical in a character, and sometimes there is no semantic radical at all. When the latter is the case, however, it is usually said that the character is a nonphonetic compound. This name is misleading since the whole character by itself represents a syllable, and thus is no less phonetic than a phonetic radical in a compound character.

[5]The misunderstanding was probably caused by the ambiguity of the description in the original paper by Tzeng and Wang (1983). I found an unambiguous description only in the caption for Figure 3 in the paper.

[6]The spellings used here are in pinyin. They are only used here to represent the Chinese characters referred to. They do not reflect the phonological values of these morphemes in Cantonese.

[7]The difference between the two studies—that is, in Woo and Hoosain, poor readers were more affected by visual similarity whereas in Ren and Mattingly good readers were more affected by phonological similarity—is probably due to the difference in presentation of the stimulus list. In the former, it was simultaneous display, whereas in the latter, it was serial display. See the discussion of the study by Mou and Anderson (1981) later in this review.

[8]Incidentally, when I was at my preliminary stage of learning English, I did not have the privilege of access to a teacher or even another fellow student, so I learned my English words by writing them over and over again. So, today, I still have great difficulty spelling words orally or understanding other people's oral spelling. But I don't have reading difficulty in English. If I became aphasic, I don't think I would be able to use spelling to help recognize English words. On the other hand, many of my fellow Chinese can do oral spelling fluently, because they started learning English with a teacher.

Appendix

SR #	Report Date	NTIS #	ERIC #
SR-21/22	January-June 1970	AD 719382	ED 044-679
SR-23	July-September 1970	AD 723586	ED 052-654
SR-24	October-December 1970	AD 727616	ED 052-653
SR-25/26	January-June 1971	AD 730013	ED 056-560
SR-27	July-September 1971	AD 749339	ED 071-533
SR-28	October-December 1971	AD 742140	ED 061-837
SR-29/30	January-June 1972	AD 750001	ED 071-484
SR-31/32	July-December 1972	AD 757954	ED 077-285
SR-33	January-March 1973	AD 762373	ED 081-263
SR-34	April-June 1973	AD 766178	ED 081-295
SR-35/36	July-December 1973	AD 774799	ED 094-444
SR-37/38	January-June 1974	AD 783548	ED 094-445
SR-39/40	July-December 1974	AD A007342	ED 102-633
SR-41	January-March 1975	AD A013325	ED 109-722
SR-42/43	April-September 1975	AD A018369	ED 117-770
SR-44	October-December 1975	AD A023059	ED 119-273
SR-45/46	January-June 1976	AD A026196	ED 123-678
SR-47	July-September 1976	AD A031789	ED 128-870
SR-48	October-December 1976	AD A036735	ED 135-028
SR-49	January-March 1977	AD A041460	ED 141-864
SR-50	April-June 1977	AD A044820	ED 144-138
SR-51/52	July-December 1977	AD A049215	ED 147-892
SR-53	January-March 1978	AD A055853	ED 155-760
SR-54	April-June 1978	AD A067070	ED 161-096
SR-55/56	July-December 1978	AD A065575	ED 166-757
SR-57	January-March 1979	AD A083179	ED 170-823
SR-58	April-June 1979	AD A077663	ED 178-967
SR-59/60	July-December 1979	AD A082034	ED 181-525
SR-61	January-March 1980	AD A085320	ED 185-636
SR-62	April-June 1980	AD A095062	ED 196-099
SR-63/64	July-December 1980	AD A095860	ED 197-416
SR-65	January-March 1981	AD A099958	ED 201-022
SR-66	April-June 1981	AD A105090	ED 206-038
SR-67/68	July-December 1981	AD A111385	ED 212-010
SR-69	January-March 1982	AD A120819	ED 214-226
SR-70	April-June 1982	AD A119426	ED 219-834
SR-71/72	July-December 1982	AD A124596	ED 225-212
SR-73	January-March 1983	AD A129713	ED 229-816
SR-74/75	April-September 1983	AD A136416	ED 236-753
SR-76	October-December 1983	AD A140176	ED 241-973
SR-77/78	January-June 1984	AD A145585	ED 247-626
SR-79/80	July-December 1984	AD A151035	ED 252-907
SR-81	January-March 1985	AD A156294	ED 257-159
SR-82/83	April-September 1985	AD A165084	ED 266-508
SR-84	October-December 1985	AD A168819	ED 270-831
SR-85	January-March 1986	AD A173677	ED 274-022
SR-86/87	April-September 1986	AD A176816	ED 278-066
SR-88	October-December 1986	PB 88-244256	ED 282-278

SR-89/90	January-June 1987	PB 88-244314	ED 285-228
SR-91	July-September 1987	AD A192081	**
SR-92	October-December 1987	PB 88-246798	**
SR-93/94	January-June 1988	PB 89-108765	**
SR-95/96	July-December 1988	PB 89-155329	**
SR-97/98	January-July 1989	PB 90-121161	ED32-1317
SR-99/100	July-December 1989	PB 90-226143	ED32-1318
SR-101/102	January-June 1990	PB 91-138479	ED325-897
SR-103/104	July-December 1990	PB 91-172924	ED331-100
SR-105/106	January-June 1991	PB 92-105204	ED340-053
SR-107/108	July-December 1991	PB 92-160522	ED344-259
SR-109/110	January-June 1992	PB 93-142099	ED352594
SR-111/112	July-December 1992	PB 93-216018	ED359575

AD numbers may be ordered from:

U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Road
Springfield, VA 22151

ED numbers may be ordered from:

ERIC Document Reproduction Service
Computer Microfilm Corporation (CMC)
3900 Wheeler Avenue
Alexandria, VA 22304-5110

In addition, Haskins Laboratories Status Report on Speech Research is abstracted in *Language and Language Behavior Abstracts*, P.O. Box 22206, San Diego, CA 92122

**Accession number not yet assigned

Haskins
Laboratories
Status Report on

SR-113
JANUARY-MARCH 1993

Speech Research

Contents

- Some Assumptions about Speech and How They Changed
  Alvin M. Liberman 1
- On the Intonation of Sinusoidal Sentences: Contour and Pitch Height
  Robert E. Remez and Philip E. Rubin 33
- The Acquisition of Prosody: Evidence from French- and English-Learning Infants
  Andrea G. Levitt 41
- Dynamics and Articulatory Phonology
  Catherine P. Browman and Louis Goldstein 51
- Some Organizational Characteristics of Speech Movement Control
  Vincent L. Gracco 63
- The Quasi-steady Approximation in Speech Production
  Richard S. McGowan 91
- Implementing a Genetic Algorithm to Recover Task-dynamic Parameters of an Articulatory Speech Synthesizer
  Richard S. McGowan 95
- An MRI-based Study of Pharyngeal Volume Contrasts in Akan
  Mark K. Tiede 107
- Thai
  M. R. Kalaya Tingsabadh and S. Arthur Abramson 131
- On the Relations between Learning to Spell and Learning to Read
  Donald Shankweiler and Eric Lundquist 135
- Word Superiority in Chinese
  Ignatius G. Mattingly and Yi Xu 145
- Prelexical and Postlexical Strategies in Reading: Evidence from a Deep and a Shallow Orthography
  Ram Frost 153
- Relational Invariance of Expressive Microstructure across Global Tempo Changes in Music Performance: An Exploratory Study
  Bruno H. Repp 171
- A Review of *Psycholinguistic Implications for Linguistic Relativity: A Case Study of Chinese* by Rumjahn Hoosain
  Yi Xu 197

Appendix 209